126 P

UNPUBLISHED PRELIMINARY DATA

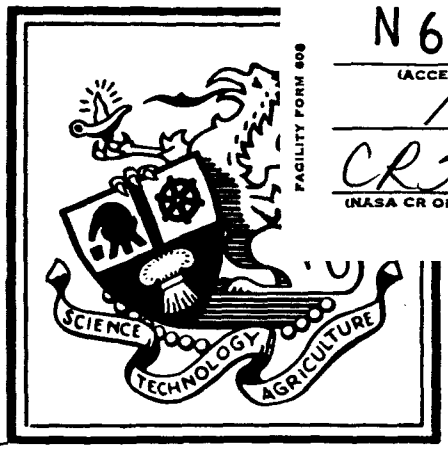
PURDUE UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING
ELECTRONIC SYSTEMS RESEARCH LABORATORY

ON A CLASS OF APPROXIMATION PROBLEMS IN SIGNAL DESIGN

by

J. C. Hancock and R. E. Totty

Technical Report No. TR-EE 64-13
Supported by National Aeronautics
and Space Administration under
Grant-NsG-553



AUGUST 1964
LAFAYETTE, INDIANA

RESEARCH GRANT

No. NsG-553

PRF 3823

ON A CLASS OF APPROXIMATION PROBLEMS IN SIGNAL DESIGN

for

NATIONAL AERONAUTICS AND SPACE

ADMINISTRATION

WASHINGTON 25, D. C.

by

J.C. Hancock, Principal Investigator

R.E. Totty

School of Electrical Engineering

Purdue University

Lafayette, Indiana

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

ABSTRACT

All communication systems are subject ot approximation error. An inevitable source of error is the difference between the realized signals and filters and the intended ones. In general, quantative analysis of the effect of these discrepencies is quite difficult, and even the selection of an error criterion that is both physically meaningful and mathematically tractable may be a problem in itself. However, for a large class of receivers, the so called correlation detectors or matched filters, an error analysis free of the signal and filter detail can be carried out. The nature of these receivers makes the choice of the mathematically tractable integral-squared-error criterion a physically meaningful one. Some of the factors which may cause the actual performance of matched filter receivers to be less than their expected performance are: (1) The matched filters may deviate from their intended design or the transmitted signals may be different from their designed waveforms; (2) The channel may alter the waveforms of the transmitted signals; (3) If a discrete receiver is used, it may not be a sufficiently good approximation of the analog receiver.

The effect of these discrepencies on probability of error for three binary systems is considered. Bounds are obtained on degradation of performance of On-Off, Antipodal, and Orthogonal Signal Systems due to mismatching of the signals or the filters. It is shown that the Equal-Energy-Orthogonal-Signals System is potentially the most sensitive of the three systems considered, and that for this case, mismatching of the signals is more serious than the same amount of mismatch in the filters.

In an attempt to compensate for the errors due to (2), procedures are developed which allow the transmitted signals to be "pre-distorted" in order that the received signals have a desired relationship (e.g. the transmitted signals might be constructed so that the received signals are orthogonal).

Examination of the error arising from (3) brings to light some subtleties concerning the discrete receiver, particularly the concepts of finite-dimensional signal representation. In connection with this latter problem we derive some useful and computationally simple expressions for the approximation error incurred in approximating a countable sum of exponentials by an element of the subspace spanned by a finite number of other exponentials.

Chapter 1

INTRODUCTION

1.1 Signal Design in Communication Systems

It is probably true that, as stated by Brennan [9], "in the last analysis, communication systems are designed by 'seat-of-the-pants' engineering let no forest of formulae --- suggest otherwise." The role of "Communication Theory" has been one of providing upper and lower bounds on the performance of communication systems. By analysis based on more or less idealized models of real communication systems, limits on performance has been obtained, thus by-passing profitless experiments and innovations. The fact that long established "seat-of-the-pants-engineered" systems are already operating very close to attainable limits does not detract from the results of communication theory in providing these limits.

The above quotation concerning design of real communication systems is certainly revelent to that part of communication system design concerning the information-bearing signals. Nonetheless, studies of the properties of signals, their selection, optimization, realization, etc., are important and useful in their own right, although immediate application in this or that communication system may not be evident. A study of systems without signals and viceversa is neither meaningful nor feasible. There is invariably a joint concern for the signals and systems. A collection of papers dealing with signal theory in relation to system theory (not necessarily communication system theory) may be found in [34].

In digital communication systems, application of statistical decision theory has proven to be of value not only in providing limits of performance,

but also in suggesting ways by which better digital systems may be built. The statistical decision theory approach is not without its limitations, and it is certainly no panacea for communication problems. Not the least of its limitations is the functional complexity involved when the signals are corrupted by other than additive gaussian noise. The "information" is conveyed in a digital communication by the decision as to which one of several possible signals has been transmitted. By assigning costs to various types of errors, a receiver structure which minimizes the average cost may, in some cases, be found. In an "optimum" receiver structure can be derived for arbitrary signals, the average cost may, in principle, be further reduced by proper choice of the transmitted signals. The "proper choice" of the transmitted signals is influenced by the channel, or media, through which the signals propagate, and the properties of the noise corrupting the signals. With perhaps one exception [35], minimum cost or Bayes receivers are developed on the assumption that the received signal is neither preceded nor followed by any other signal. That is, the effect of signal-overlap is either ignored or is assumed to have little effect on the system. The selection of optimum signals to enhance the performance of a Bayes receiver is, of course, precluded when the form of the receiver is not known (eq. for non-gaussian noise, general multiplicative disturbances, overlapping signals etc.). Even when the optimization procedure may be successfully carried out and the waveforms of the optimum signals explicitly presented, practical considerations may prohibit actual generation of the optimum signals. At any rate, the performance of "more practical" signals be compared to that which might be obtained using the optimum signals.

1.2 Approximation and Error

All communication systems are subject to approximation error whether it comes about from the inevitable differences in the actual signals, filters,

etc. and the intended ones, or more basically, the difference in the actual system and the model on which the analysis was based. In general, quantitative analysis of the effect of these discrepencies is quite difficult, and even the selection of an error criterion that is both physically meaningful and mathematically tractable may be a problem in itself. In general, different communication systems with their given signals, channels, and receivers are subject to different analysis as regards their sensitivity to changes or derivations in their signals, filters, etc. However, for a large class of receivers, the so-called correlation detectors or matched filter, an error analysis free of the signal and filter detail can be carried out. Under the assumptions of fixed signals and additive gaussian noise, the receiver structure obtained from the Bayes formulation takes the form of a correlation detector. The decision as to which signal was transmitted is based on the magnitude and sign of a statistic obtained by correlating the input data with known signal data. The number obtained by this process is proportional to the amount of energy in the signal during the correlation time interval. The fact that for these systems the performance is affected only by the apparent loss or gain of energy during the observation time interval makes the integral-squared error criterion ideally suited for examining the sensitivity of performance to deviations in the signals and filters.

1.3 Factors which affect System Performance

The degradation of the expected performance of matched filter receivers may come about for several reasons:

(1) The channel may alter the waveforms of the transmitted signals

(2) The matched filters may deviate from their intended design

(3) The discrete receiver may not be a sufficiently good approximation of the analog receiver.

The overall system considered here may be visualized as having a channel consisting of a waveform distorting filter followed by additive gaussian noise, and a receiver which is a correlator or matched filter.

The effect of these errors on probability of error for three binary systems is considered in Chapter II where the criterion of approximation error is taken to be the integral-squared-error. In Chapter III, procedures are developed to compensate for the waveform-distortion of the channel. The transmitted signals are "pre-distorted" so that the received signals have a desired relationship. Consideration of the error arising from (3) brings to light some subtlties concerning the discrete receiver, and in particular, the concept of finite-dimensional signal representation and the choice of coordinates used in the discrete receiver.

1.4 Outline of the Thesis

Chapter II is concerned mainly with the effects of filter and signal mismatch for three important binary systems. We show that for the on-off and antipodal systems, the sensitivity of performance (probability of error) is a function of only the magnitude of the integral-squared-error between the intended signal and the actual signal. Moreover, the same degradation in performance is obtained whether the error is due to the fact that the intended signal and the actual signal are different, or whether the actual matched filters differ from the ideal matched filters. The most interesting case is that of an equal-energy orthogonal-signal system. In contrast to the above two cases, we show that the character of the approximate signals (not just the magnitude of the integral-squared-error) has considerable influence on system performance. Also, for the same magnitude of error, it is shown that the cases of "imperfect signals and ideal filters," and "imperfect filters and ideal signals" are significantly different. In particular,

filter approximation error is found to be less serious than signal approximation error. Bounds are obtained on the sensitivity of performance for a given approximation error.

The first part of Chapter II provides motivation for the latter part, and deals with the selection of optimum signals to provide maximum signal-to-noise ratio at the receiver. It is shown that even for colored noise, the optimum receiver is a matched filter when the signal is optimum thus providing justification for consideration of only matched filter receivers.

Chapter III considers one source of the mismatched error considered in Chapter II; channel distortion of the transmitted signals. If the signals are received in the presence of white gaussian noise, it is well known that the total probability of error depends only on the pair-wise correlations (or inner products) of the received signals. Computationally simple procedures are developed which provide transmitted signals having the property that the received signals have a desired inner product matrix. The functional form of the channel is not required. It is required only that the channel be linear and that the inner products of the output signals can be measured by any convenient method. However, even when the construction procedure is carried out exactly, approximation error of the type considered in Chapter II must still be considered due to the approximation of the method used to measure the inner products of the output signals. As a matter of convenience, the receiver may be of the discrete type which operates on some N-tuple representation of the signal such as its time samples. In any case, the actual signals must be analog waveforms whether the receiver is discrete or not.

In Chapter IV we examine some aspects of the problem of characterizing a time waveform by an ordered N-tuple. If the filters are approximated by a

finite linear combination of other functions (e.g. by a lumped parameter filter), the effect of the approximation error is given by the result in Chapter II. Also, if the receiver is constrained to the use of particular coordinates, or if the number of coordinates is constrained, intriging but apparently quite difficult mathematical problems are uncovered. Abstractly, the problem becomes that of finding "best" finite-dimensional representations for given classes of signals. An attempt is made to make precise the intuitive notion of a set of signals being "essentially finite-dimensional". Even for mathematically well defined sets of signals, the problem is extremely difficult. For actual signals, the problem is compounded by the inability to adequately describe the ensemble of signals. Practically, all one can do is choose a finite set of representation functions and compute the resulting approximation error. A particularly convenient set (both mathematically and physically) of representation functions are the exponential functions as has been amply demonstrated particularly by W.H. Huggins and his co-workers at Johns Hopkins University. In this connection, we develop some rather remarkably simple and computationally useful approximation-error expressions for representation by exponential functions which further extends the usefulness of the exponential functions.

Although explicit results were not obtained for the problem of finding "best" approximating functions for a prescribed class of signals, some useful bounds were obtained on the maximum error to be expected using any set of N functions to approximate any one of a set of M signals. Also we examine the effect of the number of required signal coordinates for the case of conditional maximum likelihood estimates of the signal coordinates.

Chapter 2

SIGNAL SELECTION AND MISMATCH ERROR

2.1 Introduction

In the design of communication receivers it is invariably the case that the actual receivers or filters differ from their intended design. Moreover, due to the properties of the propagation media, the received signals differ from those for which the receiver was designed. These discrepencies may lead to a significant degradation in overall system performance especially in the case where the receiver is designed to receive orthogonal signals. It is possible that by proper selection of the signals to be transmitted, the performance of the system may be improved. This is the case if the channel is such that the received waveforms differ from the transmitted waveforms. Here the transmitted waveforms are selected so that the received waveforms arrive with the largest possible energy. If the additive noise is not white, the signals may be selected so that their energy is concentrated in that portion of the frequency bound where the noise power density is lower. Explicit expressions can be developed for the optimum transmitted waveforms although actual solutions are difficult to obtain. Even if the form of the "best" waveform is obtained, practical considerations may prohibit actual construction, and other signals which are easier to generate accurately may be sought whose performance is "close" to that of the optimum signal. Some examples are given in this chapter where explicit solutions for optimum waveforms are obtained and their performance compared

with other signals chosen on the basis of their ease of generation. The degree to which system performance is degraded due to the deviations of the actual signals and filters from the intended cnes is examined for binary systems utilizing three types of signaling schemes; on-off, orthogonal, and anti-podal signals. Bounds on the degradation in probability of error due to approximation errors are obtained where the receiver structure has the form of a matched filter. Since the optimum receiver is a matched filter for the case where the noise is white, and it is easily shown that for the case where the noise is colored, the optimum receiver is also a matched filter if the optimum signals are used, it suffices to consider the problem of finding the degradation in performance due to the deviation of the actual signals from the intended signals and the deviation of the actual filters from the intended matched filters.

2.2 Correlation Detectors and Matched Filters

A filter with impulse response $h(t)$ is said to be matched to a signal $x(t)$ if $h(t) = K\, x(\alpha - t)$ where K and α are arbitrary real constants. If a signal $y(t)$ is applied to a filter whose impulse response is given by $h(t) = x(T-t)$, the filter output $y(t)$ is given by

$$y(t) = \int_0^t y(\tau)\, h(t-\tau)\, d\tau$$

$$= \int_0^t y(\tau)\, x(T-t-\tau)\, d\tau .$$

The value of $y(t)$ at time T is then given by

$$y(T) = \int_0^T y(\tau)\, x(\tau)\, d\tau .$$

The number obtained by multiplying x(t) by y(t), integrating the product from t=o to t=T is the same as the number obtained by sampling, at t=T, the output of a filter with impulse response h(t) = x(T-t) when y(t) is the input to the filter. For this reason, the operation of correlating x(t) with y(t) (multiplying x(t) by y(t) and integrating from t=o to t=T) is said to be equivalent to passing y(t) through a filter matched to x(t).

If a linear filter with impulse response h(t) acts on an input $\omega(t) = x(t) + n(t)$ where x(t) is a known time function and n(t), the noise, is a sample function from a wide-sense stationary random process, the output signal-to-noise ratio

$$\frac{S}{N} = \frac{y^2(t)}{E[n_o^2(t)]}$$

is maximized at t = T if h(t) is the solution to the integral equation

$$\int_0^T R_{nn}(t-\beta)\, h(\beta)d\beta = x(T-t) \qquad 0 \leq t \leq T$$

If the noise is white, $R_{nn}(t-\beta) = \delta(t-\beta)$, and h(t) = x(T-t). That is, the filter is matched to x(t). The signal component of the output, y(t), evaluated at t = T has the value

$$y(T) = \int_0^T x^2(t)dt = E_T$$

called the energy[1] in x(t) over the time interval (0,T).

2.3 Optimum Waveforms for Signals in Colored Noise

The optimum (minimum probability of error) receiver structure for deciding whether "signal plus noise" or "noise only" was present during the observation interval (0,T) is a matched filter if the noise is white

and gaussian [1].The signal component of the output of the matched filter at t = T is given by $E_T = \int_0^T x^2(t)$. The expected value of the output noise component at t = T is given by $N_o E$ where N_o is the spectral density of the noise. For any signal x(t) with energy[1] E, the value of the output of the matched filter at t = T is the same. The choice of signal waveform, then, does not influence the system performance so long as the filter is matched to the signal and the background noise is white. If the background noise is colored, however, one can find signals whose waveforms are perferable to others as is shown by Middleton [2]. Here we derive the same result and use it to show that the optimum filter in this case is also a matched filter if the signal is the optimum signal. That is, the optimum filter is simply a filter matched to the signal which produces the maximum signal-to-noise ratio.

Suppose the additive gaussian noise has an autocorrelation function $R_{nn}(\tau)$. The optimum receiver structure for deciding "signal plus noise" or "noise only" is a filter with impulse response h(t) where h(t) satisfies the integral equation [1]

$$\int_0^T R_{nn}(t-\tau)\, h(\tau)d\tau = x(T-t). \qquad (2\text{-}1)$$

Since the input noise is gaussian, the sample at t = T of the output of the filter is a gaussian random variable whose mean is the output signal component $\mu_o(T)$. The probability of error may be decreased further by choosing the input signal x(t) so as to maximize the signal-to-noise ratio defined by

$$\frac{S}{N} = \frac{y_o^2(T)}{E[n_o^2]}$$

---

(1) Although the term "energy" is widely used to denote the integral of the square of a time function, it is best to point out here that it does not represent the energy supplied to the filter. The energy supplied to the filter is proportional to the integral of the square of the input signal only if the input impedance is purely resistive.

where n_o is the noise component of the output of the filter. The expected value of the square of the output noise sample at t = T may be written as

$$E[n_o^2] = E\{\int_0^T n(\tau)\, x(T-\tau)\, d\tau\}^2 = \int_0^T \int_0^T R_{nn}(\tau-\beta)\, h(\tau)\, h(\beta)\, d\tau d\beta. \tag{2-2}$$

The square of the output signal component at t = T is given by

$$y_o^2(T) = [\int_0^T h(\tau)\, x(T-\tau)\, d\tau]^2 \tag{2-3}$$

substituting eq.(2-1) into eq.(2-2), we have

$$\int_0^T x(T-\beta)\, h(\beta)\, d\beta \tag{2-4}$$

Dividing eq.(2-3) by eq.(2-4) yields

$$\frac{S}{N} = \int_0^T h(\tau)\, x(T-\tau)\, d\tau \tag{2-5}$$

but by the schwortz inequality, (2-5) is maximum when

$$h(\tau) = x(T-\tau)$$

From eq.(2-1) we have that

$$\int_0^T R_{nn}(\beta-\tau)\, x(T-\beta)\, d\beta = \frac{\alpha}{K}\, x(T-\beta) \qquad 0 \le \beta \le T \tag{2-6}$$

This equation has solution when $\frac{\alpha}{K} = \lambda_j$, j = 1, 2, ... where λ_j are the eigenvalues of (2-6). Using the j^{th} eigenfunction as a signal, it is seen that the optimum filter is still a matched filter, producing a S/N of

$$S/N = \frac{K}{\alpha} \int_0^T s^2(T-\beta)\, d\beta = E\, \lambda_j \tag{2-7}$$

We seek then the eigenfunction corresponding to the largest eigenvalue. Of course the eigenvalues may become arbitrarily large (i.e. there exists no largest eigenvalue) indicating that without further restrictions, one can do arbitrarily well in colored noise. This is always the case if the noise has become colored by passing through a physically realizable filter. This is, of course, intuitively obvious if, say the noise spectrum falls off as $0(\frac{1}{\omega^2})$, one would place the signal at a high enough frequency so that the level of the noise is negligible. The above is true if no bandwidth constraints are placed on the signal. It seeems to be extremely difficult to place such constraints in a variational problem of this type.

2.4 Optimum Waveforms for Channels with Impulse Response h(t)-

It is implicit in the above discussion that the received waveform has the same waveform as the transmitted waveform. In the following, it is assumed that the transmitted waveform x(t) has passed through a filter with impulse response h(t) where the output y(t) is corrupted by additive stationary white noise (Fig. 2-1)

Channel With Impulse Response h(t).

Fig. 2-1

As has been previously demonstrated, there is no preferred signal waveform when the additive noise is white; all signals with equal energy are equally desirable. It follows then that x should be selected so as to maximize the energy in y (with an energy constaint on x) during the observation time T.

The 1950 paper by Chalk [4] provides a partial solution to this problem. This paper was primarily concerned with the interference problem and essentially attempts to find the pulse type waveform which maximizes the energy received in the whole interval $(0, \infty)$. The extension to the case above is easily done although a time domain approach seems to yield the answer in a more straight forward manner.

Specifically, let the channel be represented by its impulse response $h(t)$. We wish to find the input pulse $x(t)$ having unit energy which is non-zero only on $(0, T)$ which maximizes the energy received in $(0, T)$ at the output. Now

$$y(t) = \int_0^t h(t-\alpha)\, x(\alpha) d\alpha = \int_0^T h(t-\alpha)\, \mu(t-\alpha)\, x(\alpha)\, d\alpha \qquad (2\text{-}8)$$

where the right hand side of (8) expresses the fact that $h(t)$ is realizable $(h(t) = 0\ t < 0)$ and also allows use of definite limits. μ is the unit step function defined as $\mu(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$. Then writing the square of (8) as an iterated integral we have

$$y^2(t) = \int_0^T \int_0^T \int_0^T h(t-\alpha)\, h(t-\beta)\, \mu(t-\alpha)\, \mu(t-\beta)\, x(\alpha)\, x(\beta)\, d\alpha\, d\beta \qquad (2\text{-}9)$$

The output energy $E_o = \int_0^T y^2(t)\, dt$ is given by

$$E_o = \int_0^T \int_0^T \int_0^T h(t-\alpha)\, h(t-\beta)\, \mu(t-\alpha)\, \mu(t-\beta)\, x(\alpha)\, x(\beta)\, d\alpha\, d\beta\, dt \qquad (2\text{-}10)$$

Define

$$\int_0^T h(t-\alpha)\, h(t-\beta)\, \mu(t-\alpha)\, \mu(t-\beta)\, dt = H(\alpha, \beta) \qquad (2\text{-}11)$$

Then from (2-10) we have

$$E_o = \int_0^T \int_0^T H(\alpha, \beta)\, x(\alpha)\, x(\beta)\, d\alpha\, d\beta \qquad (2\text{-}12)$$

We wish to maximize (2-12) subject to the constraint that $\int_0^T x^2(\beta)\, d\beta = 1$. The solution may be obtained directly via a theorem [3] that the maximum of (2-12) is obtained when $x(\alpha)$ is an eigenfunction of the integral equation

$$\lambda x(\alpha) = \int_0^T H(\alpha, \beta)\, x(\beta)\, d\beta \qquad (2\text{-}13)$$

corresponding to the largest eigenvalue. The maximum of (2-12) is equal to the largest eigenvalue of (2-13).

From (2-11) it is seen that $H(\alpha, \beta) = H(\beta, \alpha)$, i.e. H is symmetric. The optimum signals then enjoy the special properties of eigenfunctions of integral equations with symmetric kernels [3]. The eigenfunctions are orthogonal, the eigenvalues are real and (by virtue of (2-10) positive. Moreover the output signals due to these eigenfunctions are also orthogonal over (0, T). If $x(t)$ is the j^{th} eigenfunction of (2-13) with $E_T = 1$, then the response to $x(t)$ has energy λ; some of these properties will be made use of in Chapter IV.

We note here that maximizing the energy in (0, T) at the output does not imply that the energy outside T is minimized. Considerable interpulse interference may result if the input pulses are transmitted every T seconds.

An interesting property of these optimum signals is that in order to maximize the output energy in (0, T) for a realizable h(t), the input must be driven to zero at t = T. This is seen by noting from (2-11) that $H(T, \beta) = 0$, then from (2-13), $x(T) = 0$.

We now consider an example using a simple RC circuit, where the form of the optimum signals may be obtained.

Example: Let $h(t) = b\epsilon^{-bt}\,\mu(t)$, which is the impulse response of a lowpass RC circuit. Then

$$H(\alpha,\ \beta) = b^2 \int_0^T h(t-\alpha)\ h(t-\beta)\ \mu(t-\alpha)\ \mu(t-\beta)\ dt \tag{2-14}$$

$$= b^2 \int_\beta^T \epsilon^{-2bt}\ \epsilon^{b(\alpha+\beta)} dt \qquad \beta > \alpha$$

$$b^2 \int_\alpha^T \epsilon^{-2bt}\ \epsilon^{b(\alpha+\beta)} dt \qquad \beta < \alpha \tag{2-15}$$

$$= \frac{b}{2}\left[\epsilon^{-2b\beta} - \epsilon^{-2bT}\right]\epsilon^{b(\alpha+\beta)} \qquad \beta > \alpha$$

$$\frac{b}{2}\left[\epsilon^{-2b\alpha} - \epsilon^{-2bT}\right]\epsilon^{b(\alpha+\beta)} \qquad \beta < \alpha \tag{2-16}$$

We wish to solve $\lambda\,\phi(\alpha) = \int_0^T H(\alpha,\ \beta)\ \phi(\beta)\ d\beta$

$$\lambda\phi(\alpha) = \int_0^\alpha H(\alpha,\ \beta)\ \phi(\beta)\ d\beta + \int_\alpha^T H(\alpha,\ \beta)\ \phi(\beta)\ d\beta\ . \tag{2-17}$$

$\beta < \alpha$ $\qquad\qquad$ $\beta > \alpha$

Substituting (2-16) into (2-17) we have

$$\frac{2}{b}\lambda\phi(\alpha) = \epsilon^{-b\alpha}\int_0^\alpha \epsilon^{b\beta}\,\phi(\beta)\,d\beta - \epsilon^{b\alpha}\,\epsilon^{-2bT}\int_0^\alpha \epsilon^{b\beta}\,\phi(\beta)\,d\beta$$

$$-\epsilon^{+b\alpha}\int_T^\alpha \epsilon^{-b\beta}\,\phi(\beta)\,d\beta + \epsilon^{b\alpha}\,\epsilon^{-2bT}\int_T^\alpha \epsilon^{b\beta}\,\phi(\beta)\,d\beta \tag{2-18}$$

Differentiating (2-18) with respect to α, we have

$$\frac{2}{b^2}\lambda\dot{\phi}(\alpha) = -\epsilon^{-b\alpha}\int_0^\alpha \epsilon^{b\beta}\,\phi(\beta)\,d\beta - \epsilon^{-b\alpha}\,\epsilon^{-2bT}\int_0^\alpha \epsilon^{b\beta}\,\phi(\beta)\,d\beta$$

$$-\epsilon^{b\alpha}\int_T^\alpha \epsilon^{-b\beta}\,\phi(\beta)\,d\beta + \epsilon^{b\alpha}\,\epsilon^{-2bT}\int_T^\alpha \epsilon^{b\beta}\,\phi(\beta)\,d\beta\ . \tag{2-19}$$

Subtracting (2-19) from (2-18) yields

$$\frac{2}{b}\lambda\phi(\alpha) - \frac{2}{b^2}\lambda\dot{\phi}(\alpha) = 2\,\epsilon^{-b\alpha}\int_0^{\alpha}\epsilon^{b\beta}\,\phi(\beta)\,d\beta\,. \tag{2-20}$$

Differentiating (2-20), we have

$$\frac{2}{b}\lambda\dot{\phi}(\alpha) - \frac{2}{b^2}\lambda\ddot{\phi}(\alpha) = -2b\,\epsilon^{-b\alpha}\int_0^{\alpha}\epsilon^{b\beta}\,\phi(\beta)\,d\beta + 2\,\phi(\alpha). \tag{2-21}$$

Multiplying (2-20) by b and adding (2-21) yields finally

$$\ddot{\phi}(\alpha) + \frac{1-\lambda}{\lambda}\,b^2\,\phi(\alpha) = 0 \tag{2-22}$$

The solutions of (2-17) must then satisfy (2-22). Let $\delta^2 = \frac{1-\lambda}{\lambda}$.

The solution of (2-22) has the form

$$\phi(t) = A\cos\delta bt + B\sin\delta bt. \tag{2-23}$$

We have additional information that, as was noted earlier, $\phi(T) = 0$. This requires that

$$\tan\delta bT = -A/B \tag{2-24}$$

rewriting (2-23), we have

$$\phi(t) = B\,[-\mathrm{Tan}\,\delta bT\cos\delta bt + \sin\delta bt]. \tag{2-25}$$

Since the solution of (2-17) is independent of B, we replace it by unity. With λ replaced by $\frac{1}{1+\delta^2}$ in (2-17), (2-25) is substituted into (2-17) to determine δ or δ's that satisfy the integral equation (2-17). It turns out that (2-25) satisfies (2-17) if

$$\mathrm{Tan}\,\delta bT = -\delta \tag{2-26}$$

Verification that this is true is straightforward but is rather tedious and is omitted here.

We solve (2-26) for the case bT = 1. By trial and error we find the smallest non-zero solution of (2-26) to be

$$2.029 < \delta < 2.030$$

and $0.1952 < \lambda < 0.1954$.

That is, the maximum ratio of $\dfrac{\int_0^T y^2(t)\,dt}{\int_0^T x^2(t)\,dt}$ for this system is about 0.195 (for bT = 1), and is attained with input signals of the form (2-25), where δ is given by (2-26).

Differential Equation Formulation

For illustrative purposes, the problem is now formulated utilizing the differential equation of the simple RC circuit. The input x is related to the output y by the differential equation

$$\dot{y} + by = bx \tag{2-27}$$

we now seek to maximize $\int_0^T y^2(t)\,dt$ while constraining $\int_0^T x^2(t)\,dt$ to say unity. Equivalently, we seek to extremize the integral

$$I = \int_0^T \left[y^2(t) - \lambda \left(\frac{1}{b}\, y(t) + \dot{y}(t) \right)^2 \right] dt \tag{2-28}$$

$$= \int_0^T f(t,y,\dot{y})\,dt \tag{2-29}$$

For which the Euler equation [5]

$$\frac{\partial f}{\partial y} - \frac{d}{dt}\left[\frac{\partial f}{\partial \dot{y}} \right] = 0 \tag{2-30}$$

becomes

$$\ddot{y} + \frac{1 - \lambda}{\lambda} b^2 y = 0 \tag{2-31}$$

Two end conditions are needed here. One is found by noting that for square integrable inputs, the output is zero at t = 0, i.e. y(o) = 0. The other comes from the undetermined right end point condition

$$\frac{\partial f}{\partial y}\bigg|_{t = T} = 0 \tag{2-32}$$

or

$$\frac{1}{b}\dot{y} + y\bigg|_{t = T} = 0 \tag{2-33}$$

but $x = \frac{1}{b}\dot{y} + y$ so (2-33) implies x(T) = 0, agreeing with the general condition found earlier that the optimum signal for any realizable filter drives the input to zero at t = T. The general solution of (2-31) (for $0 < \lambda < 1$) is given by

$$y(t) = A \cos \delta bt + B \sin \delta bt \tag{2-34}$$

where $\frac{1 - \lambda}{\lambda} = \delta^2$. Imposing the above endpoint conditions yields

$$y(t) = B \sin \delta bt \tag{2-35}$$

where δ is given by

$$\text{Tan } \delta b T = - \delta \tag{2-36}$$

This solution agrees with the previous solution as it should, and was obtained with a good deal less effort. Although the formulation of the problem in terms of the impulse response of the filter is more general, explicit solutions of equations such as (2-13) with finite limits are known only for a few special types of kernels. Since if the filter is such that

the input and output are related by an ordinary differential equation, the solution of (2-13) must reduce to the solution of an Euler equation derived using the differential equation, it is more direct to use the differential equation approach. Moreover, there seeems to be no way of including endpoint constraints using the integral equation approach. For systems of higher order one encounters the well known difficulties of solving differential equations with conditions specified at each endpoint, leading to solution of sets of transcendental equations. However, the actual waveforms are not as important as finding the largest possible values $Q = \int_0^T y^2(t)\,dt \,/\, \int_0^T x^2(t)\,dt$ and comparing this value with the performance obtained from signals that are more easily and accurately generated. For example, if we select $x(t) = \epsilon^{-\nu t}$ the system is being driven at its natural frequency, and we would expect fairly good performance. For $bT = 1$ we find that $Q = 0.97\, Q_{max}$.

We note that input signals of the form (2-26), $x(t) = \delta \cos \delta bt + \sin \delta bt$, produce output signals having no transient terms. In the frequency domain the input signal is such that it contains a zero where the filter has a pole.

In the above formulations of the problem of maximizing Q, no consideration was given to the behavior of $y(t)$ for $t > T$. The other extreme of requiring $y(t) = 0$ for $t > T$ may be considered using the method of Diamond and Gerst [6] for filters whose input and output are related by an ordinary differential equation. The penalty for requiring $y(t) = 0$ $t > T$ is a decrease in Q_{max}. A detailed investigation of this case is presently being conducted by H. Schwarzlander [33] at Purdue University.

When the output is constrained to be a pulse the solution for the RC circuit is given in [17](2). For bT = 1, the optimum pulse producing input requires roughly twice as much "Energy" as does the optimum input signal given by (2-25) (for the same output "energy" in (0, T)).

2.5 Approximation Errors

Inevitably, approximation errors arise when an attempt is made to synthesize a matched filter for a given signal. On the other hand the filter may be quite closely matched to the intended signal, but dispersion of the channel, imperfect syncronization, etc. the actual received signal may differ from the intended signal. The filter synthesis problem may play a large role in selecting the transmitted signals i.e. select those signals whose matched filters may be more easily and accurately built.

Here we examine the effect of approximation error on probability of error for some particular binary detection systems.

We first introduce some notation that is used neither for elegance nor generality but is simply less cumbersome and is also easier to type. Let $\int_T x(t)\, y(t)\, dt = (x, y)$ and call this number the "inner product of x and "y". Then let $(x, x)^{1/2} = ||\, x\, ||$ which is called the "norm" of x.

2 6 The"On-Off" Case

We consider now the problem of deciding whether a known signal x is present along with white gaussian noise, or whether noise alone is present, i.e. the "on-off" case. The form of the optimum receiver is well known

---

(2) There is an error here; the right hand sides of Eqs. (10) and (11) in [7] should be multiplied by RC.

and is shown in Fig. (2-2).

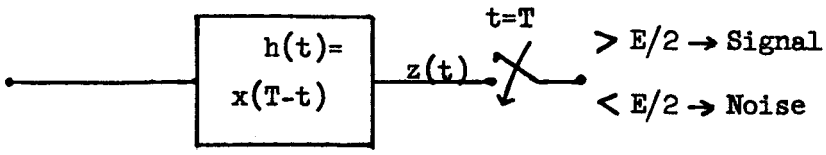


Fig. 2-2

Optimum Receiver for On-Off Case

In all the cases that are to follow, the two signals are considered to have equal a priori probabilities of occurance and that the cost of mistaking either signal for the other is the same. In the "on-off" case, with signal "energy" E, this set of assumptions results in the receiver announcing "signal plus noise" if $z(T) > \frac{E}{2}$, and "noise only" if $z(T) < \frac{E}{2}$, where $\frac{E}{2}$ is called the threshold. In the two other cases to follow (equal energy orthogonal signals, and anti-podal signals) the threshold is zero. We examine the effects of "mismatch" of the signal and the filter. If the selected signal is x_*, the actual signal used may differ somewhat from x_*, and moreover the matched filter may also differ from the intended signal x_* We consider here that either the filter is exact, and the signal is in error, or vice versa.

The criterion of error here is taken to be the normalized square error

$$\epsilon^2 = \frac{||x - x_*||^2}{||x|| \quad ||x_*||} \qquad (2\text{-}37)$$

In order to insure that the error is due to mismatch and not amplitude difference, we set $||x||^2 = ||x_*||^2 = E$ where E is the "energy" in the signal. The signal component of the output of the matched filter is E

if the signal is present and the filter is perfectly matched to the signal. If the filter is matched to x_* and signal x is sent, the output of the matched filter is given by (x, x_*). Rewriting (2-37),

$$\epsilon^2 = \frac{(x - x_*,\ x - x_*)}{||x||\ \ ||x_*||} = \frac{(x,x) - 2(x,x_*) + (x_*,x_*)}{||x||\ \ ||x_*||} \qquad (2\text{-}38)$$

recalling that $||x||^2 = ||x_*||^2 = E$, we have

$$(x,x_*) = E\left[1 - \frac{\epsilon^2}{2}\right] \qquad (2\text{-}39)$$

so that the output is degraded by the factor $\left[1 - \frac{\epsilon^2}{2}\right]$. The effect of this degradation is shown in Fig. (2-4) for various values of ϵ^2. We note here that all signals x satisfying (2-37) have the same effect on probability of error. This is true whether the error is in the filter or in the signal.(3)

2.7 Equal Energy Orthogonal Signals, Mismatched Signals

For the case of detecting the presence of either of two equal energy orthogonal signals, the previous statement is no longer true The optimum receiver for this case is well known and is shown in Fig. (2-3).

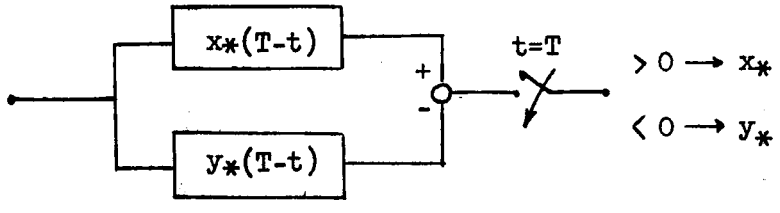


Fig. 2-3
Optimum Receiver for Orthogonal Signals

(3) For the on-off case, since the threshold is set at one half the expected signal energy, part of the degradation in probability of error shown in the curves is due to the incorrect threshold setting. For the range of mismatch error considered, this error is negligible compared to the mismatch error. For the other two cases considered, the threshold is always zero (when the signals have equal energy)

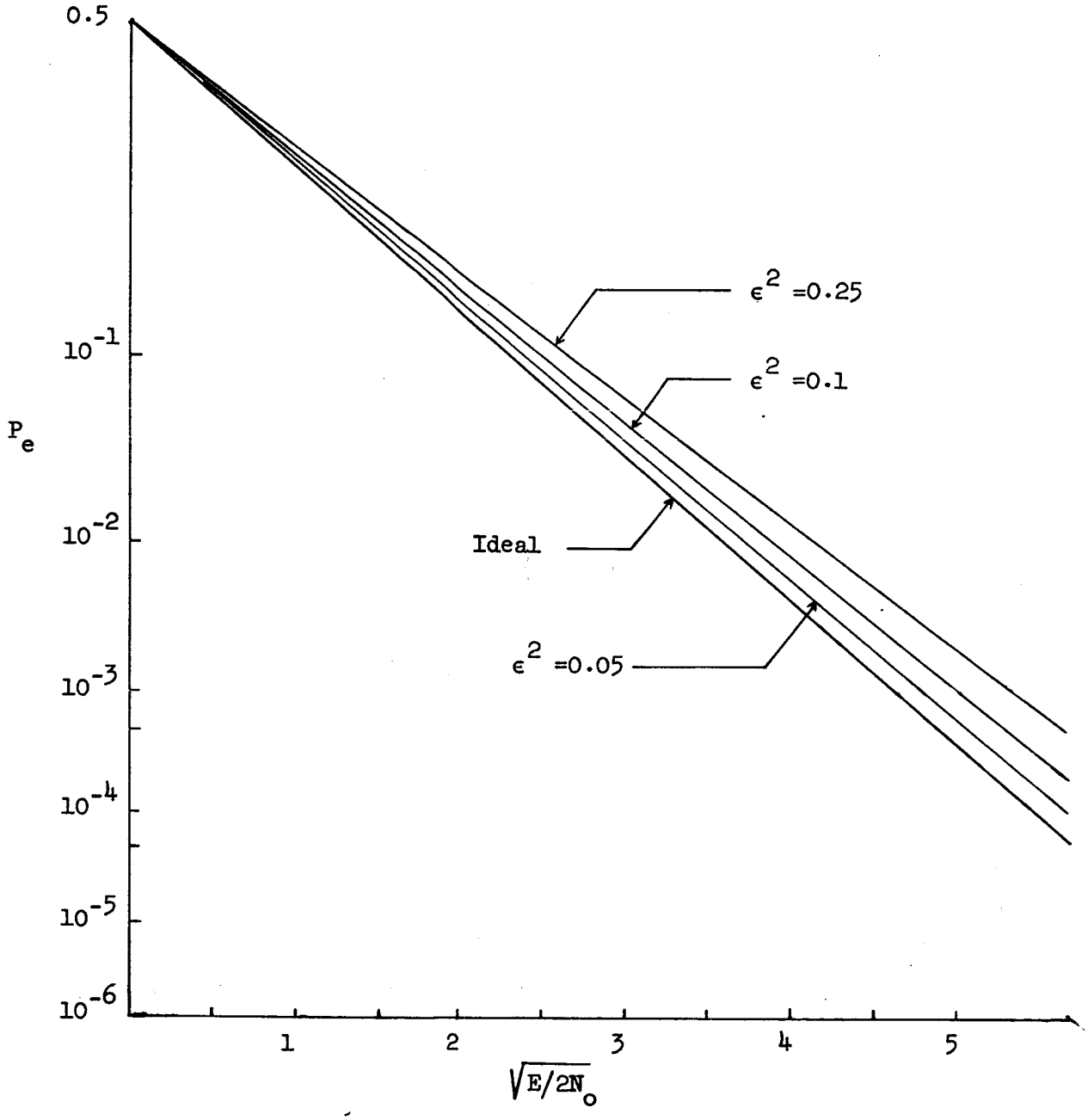


Fig. 2-4

Probability of Error for Mismatched Signals or Mismatched Filters; On-Off Case.

If the orthogonal signals x_* and y_* are matched perfectly to their filters, then the signal component of the output is E if x_* is present, and -E if y_* is present.

We now assume that the filters are perfectly matched to the intended signals x_* and y_*, but that the actual signals x and y are different from x_* and y_*; in particular to maintain symmetry, we require

$$(x,y_*) = (y,x_*) \text{ and } \frac{||x_* - x||^2}{||x|| \; ||x_*||} = \epsilon^2. \tag{2-40}$$

In this case, there are two sources of error; the error due to the diminished output of the filter matched to x_* caused by the mismatch of x and x_*, and the error due to the fact that x and y_* may no longer be orthogonal, and the output of the filter matched to y_* due to x effectively subtracts from the total output. This latter error turns out to be the more significant for the "worst case" conditions.

The degradation due to the mismatch of x and x_* is given by (2-39). The output of the filter matched to y_* due to x is given by (x,y_*) The problem then becomes: Given (2-40) and the condition that $(x,y_*) = 0$, find (x,y_*). There is, of course, no unique solution. The functions x_* and y_* are completely arbitrary except that they are orthogonal and are square integrable. The function x is constrained only by the fact that is is square integrable and is "close" to x_*. The worst case arises when (x,y_*) has the largest possible value.

We now show that if

$$||x_*|| = ||y_*|| = ||x|| = 1, \; ||x_*-x||^2 = \epsilon^2 = ||y_*-y||^2$$

and $(x_*,y_*) = 0$

then

$$(x,y_*) \leq \epsilon \sqrt{1 - \frac{\epsilon^2}{4}} \tag{2-41}$$

That is, all the signals and filters have equal energy; the actual signals (x and y) differ from the intended signals (x_* and y_*) and have the same magnitude of error; the signals x_* and y_* are orthogonal. We want to show that under these conditions, the largest possible output of the filter matched to y_* (or x_*) due to signal x (or y) is given by $(x,y_*) \leq \epsilon \sqrt{1 - \frac{\epsilon^2}{4}}$.

To show this, we first choose a set of functions $F = [f_1, f_2, \ldots]$ which is complete in the space of which x, y, x_* are members. This can always be done, since this space is L_T^2 and is known to be separable. We now form $F' = [x_*, y_*, f_1, f_2, \ldots]$ by adjoining x_* and y_* to F. F' is still complete. Form $\phi = [\phi_1, \phi_2, \phi_3, \ldots]$ by orthonormalizing the set F' where we set $\phi_1 = x_*$ and $\phi_2 = y_*$. Since Φ is complete we may expand x in terms of the ϕ's, i.e.

$$x = \sum_{i=1}^{\infty} a_i \phi_i \tag{2-42}$$

in the sense that

$$||x||^2 = \sum_{i=1}^{\infty} a_i^2 \tag{2-43}$$

and the a_i are given by $a_i = (x, \phi_i)$.

Now $||x||^2 = \sum_{i=1}^{\infty} a_i^2 = 1$. Hence

$$a_1^2 + a_2^2 \leq 1 \tag{2-44}$$

$$\text{or} \quad a_2 \leq \sqrt{1 - a_1^2} \quad . \tag{2-45}$$

Since $x_* = \phi_1$, and $y_* = \phi_2$, we have that

$$a_1 = (x,x_*) \text{ and } (y_*,x) = a_2. \tag{2-46}$$

From (2-39), $(x,x_*) = 1 - \frac{\epsilon^2}{2}$ so that

$$a_2 \leq \sqrt[+]{1 - (1 - \frac{\epsilon^2}{2})^2} = \epsilon \sqrt[+]{1 - \frac{\epsilon^2}{4}}. \tag{2-47}$$

For this "worst case", the output of the system is degraded by the factor

$$\rho = (1 - \frac{\epsilon^2}{2} - \epsilon \sqrt[+]{1 - \frac{\epsilon^2}{4}}). \tag{2-48}$$

Recall now that from Eq. (2-40), we are maintaining symmetry so that (3-38) holds for $\epsilon^2 \leq 2\sqrt{2}$, so that $0 \leq \rho \leq 1$. The dominant degrading factor in (2-48) in the second term due to the non-orthogonality of x and y^* (or y and x^*) since

$$\epsilon \sqrt{1 - \frac{\epsilon^2}{4}} > \frac{\epsilon^2}{2} \text{ for } \epsilon^2 < 2. \tag{2-49}$$

The degradation in performance of an optimum receiver designed for reception of equal energy, equally likely orthogonal signals for this worst case condition is shown in Fig.(2-5). In Fig.(2-8), the rate of decrease in performance vs. ϵ^2 is shown for a given initial probability of error.

It is interesting to compare a "poor equal energy orthogonal system with a "good" on-off system. An equal energy orthogonal signal system with $e^2 = 0.08$, for the "worst case" condition has the same performance as a perfect on-off system.

In Fig. (2-8) is also shown the rate of decrease in performance vs. ϵ^2 for an on-off system. Here we have no "worst case" conditions as all signals with the same error have the same performance.

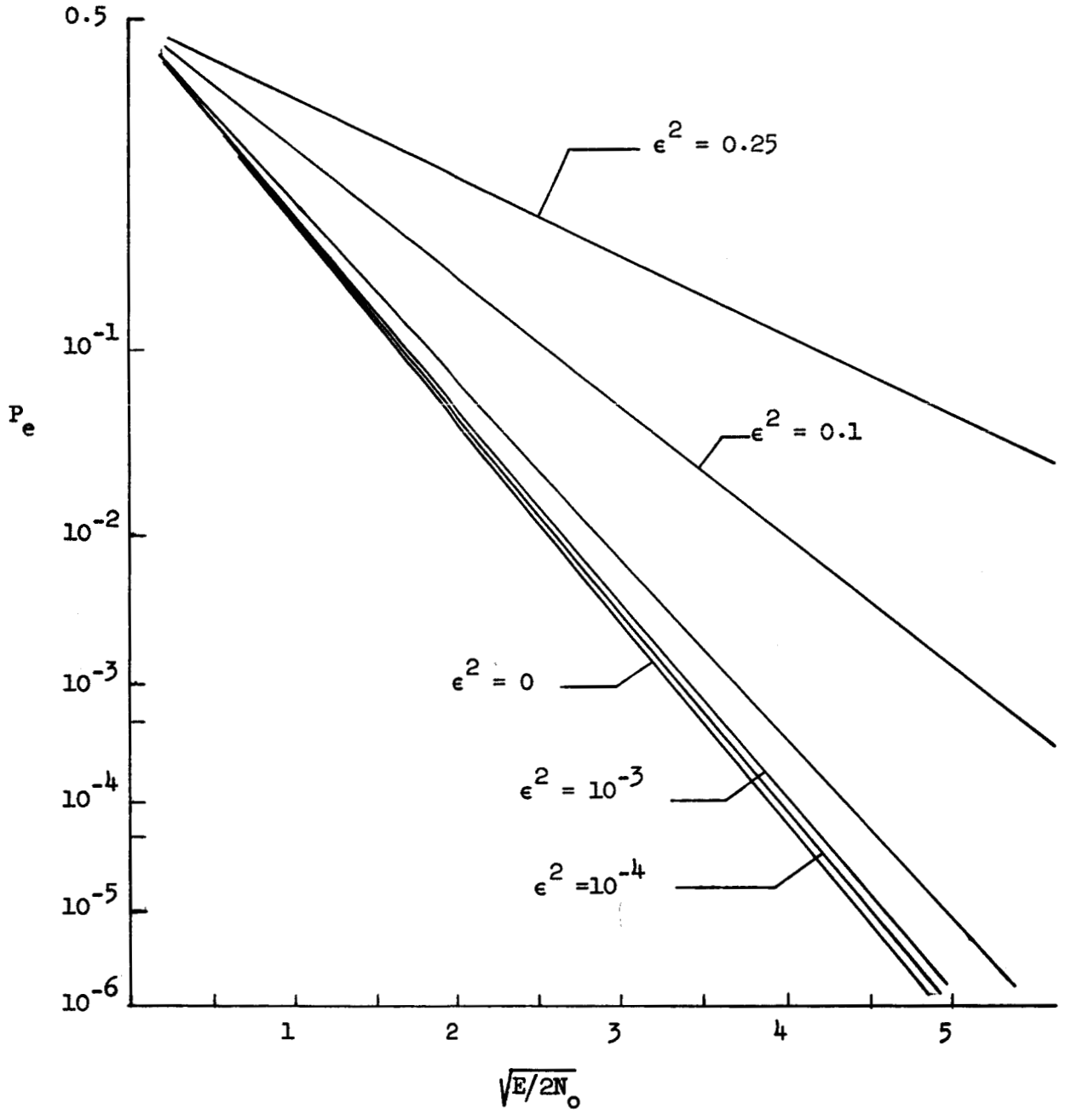


Fig. 2-5

Probability of Error for Mismatched Signals and Ideal Filters; Equal Energy Orthogonal Signals (Worst Case)

Here we are considering the case where the signals and their filters are mismatched in the same manner so that $\epsilon^2 \leq 2 - \sqrt{2}$. If $\epsilon^2 = 2 - \sqrt{2}$, $x = y$ or $x = -y$ and the factor ρ becomes zero, yielding probability of error of 0.5. For $\epsilon^2 > 2 - \sqrt{2}$ we have the situation of each signal "looking more like the other's matched filter" so that the probability of error is greater than one half.

2.8 Equal Energy Orthogonal Signals: Mismatched Filters

Although at first glance it might seem that considering perfect filters and mismatched signals is the same as the case of perfect signals and mismatched filters. This of course is not true as one would then be led to the conclusion that by building matched filters nearly orthogonal but with $(x,y_*) < 0$ and $(y,x_*) < 0$, one could improve performance (lower the probability of error). This is naturally false, as for a given signal, the matched filter receiver minimizes probability of error.

The difference is due to the variance of the noise sample at $t = T$ which is given by

$$\sigma^2 = N_o \int_0^T h^2(t)\, dt \tag{2-53}$$

where

$$h(t) = x_*(T-t) - y_*(T-t). \tag{2-54}$$

Then

$$\sigma^2 = 2N_o E\, [1 - (x_*, y_*)]. \tag{2-55}$$

If the filters are perfect, $(x_*, y_*) = 0$ and $\sigma^2 = 2EN_o$. Hence regardless of the character of the signal, the variance of the noise sample is the same as that for the ideal system, but the signal component of the sample may be increased by having (x,y_*) and (y,x_*) negative.

Considering the best possible situation where the signals are perfect and the filters are negatively "correlated", it turns out that the increase in the signal component of the sample is exactly offset by the increased variance of the noise sample. Also if the filters have equal positive correlation with the signals, and the signals have the largest positve correlation allowable, then again, the decrease in one signal component of the output sample is exactly offset by the decrease in the variance of the noise sample.

For anti-podal signals (i.e. $x = -y$), the ideal performance is better than the ideal "on-off" system (by 6 db) and is better than the ideal "orthogonal" system by 3 db. The system consists of a filter matched to x whose output is sampled at $t = T$ and if greater than zero announces x, if less than zero, -x. Here the variance of the noise remains the same, and the signal component of the sample can only be decreased by the factor $(1 - \frac{\epsilon^2}{2})$. The effect on probability of error for this case is shown in Fig. (2-7) where some of the curves for the other two cases are given for comparison. In all of the curves $0 \leq \epsilon^2 \leq 0.25$.

The curves in Fig. (2-7) showing the decrease in performance as a function of ϵ^2 for a given $S/N = \sqrt{E/2N_o}$ indicates that the binary orthogonal system is a potentially "over sensitive" system. It should be kept in mind that these curves are for the worst case.

In contrast to the on-off and anti-podal systems where only the magnitude of the error affects the performance, i.e. all signals x such that $||x-x_*||^2 = \epsilon^2$ have the same effect on probability of error, the orthogonal system is sensitive to the character of the error. The mismatch of signals may improve, degrade, or leave unaffected the probability of error. Also, for the on-off and anti-podal systems, it makes no difference whether the error is in the

filter or in the signals. This is not true for the orthogonal system.

For the orthogonal system we now examine the "worst case" for perfect signals and mismatched filters which is considerably different from the situation of perfect filters and mismatched signals. As was noted before, the difference arises in the variance of the noise in the two cases. If signal x is sent, the signal component of the output at t = T is $E[(x,x_*) - (x,y_*)]$ where E is the energy of each signal, and if signal y is sent, the signal component is $E[-(y,y_*) + (y,x_*)]$. The variance of the noise sample is in both cases given by $2EN_o[1-x_*y_*] = \sigma^2$. By straightforward substitution and change of variable, we find the probability of error is given by

$$P_e = \frac{1}{2} \int_a^\infty \epsilon^{\frac{-z2}{2}} dz \qquad (2\text{-}56)$$

$$\text{where} \quad a = \sqrt{\frac{E}{2N_o}} \; \frac{(xx_*) - (xy_*)}{\sqrt{1 - (x_*,y_*)}} ,$$

and for symmetry we require $(x,y_*) = (y,x_*)$.

Let

$$\gamma = \frac{(x,x_*) - (x,y_*)}{\sqrt{1 - (x_*,y_*)}} \qquad (2\text{-}57)$$

We now show that if $||x_*|| = ||y_*|| = ||x|| = ||y|| = 1$, $(x,y) = 0$ $||x-x_*||^2 = ||y-y_*||^2 = \epsilon^2 \leq 2 - \sqrt{2}$, and $(x,y_*) = (y,x_*)$, that

$$\sqrt{2\left[1 - \frac{\epsilon^2}{2}\right]^2} - 1 \leq \gamma \leq 1 \qquad (2\text{-}58)$$

We proceed now as before except that here $(x,y) = 0$, and we set $x = \phi_1$, $y = \phi_2$. For simplicity of notation, let $(x,x_*) = (y,y_*) = \alpha$, and $(x,y_*) = (y,x_*) = \beta$.

Then

$$x_* = \alpha\phi_1 + \beta\phi_2 + \sum_{i=3}^{\infty} a_i \phi_i \tag{2-59}$$

and

$$y_* = \beta\phi_1 + \alpha\phi_2 + \sum_{i=3}^{\infty} b_i \phi_i \tag{2-60}$$

Note that the function x_* and y_* are completely arbitrary; we know only that $(x,x_*) = (y,y_*) = \alpha$ and $(x,y_*) = (y,x_*) = \beta$. We have

$$(x_*,y_*) = 2\alpha\beta + \sum_{i=3}^{\infty} a_i b_i \,, \tag{2-61}$$

and from the Schwartz inequality,

$$\left| \sum a_i b_i \right| \leq \left(\sum_{i=3}^{\infty} a_i^2 \right)^{1/2} \left(\sum_{i=3}^{\infty} b_i^2 \right)^{1/2} \tag{2-62}$$

but $||x_*|| = ||y_*|| = 1$, so that

$$\sum_{i=3}^{\infty} a_i^2 = \sum_{i=3}^{\infty} b_i^2 = 1 - [\alpha^2 + \beta^2] \tag{2-63}$$

Then

$$(x_*,y_*) \geq 2\alpha\beta - \{1 - [\alpha^2 + \beta^2]\} \tag{2-64}$$

$$= (\alpha + \beta)^2 - 1$$

and $$(x_*,y_*) \leq 1 - (\alpha - \beta)^2$$

Substituting (2-64) into (2-57), we have

$$\frac{\alpha - \beta}{\sqrt{1 - [1-(\alpha - \beta)^2]}} = \delta(\beta) \geq \gamma \geq \phi(\beta) = \frac{\alpha - \beta}{\sqrt{2 - (\alpha + \beta)^2}} \tag{2-65}$$

but $\delta(\beta) \equiv 1$, so that probability of error cannot be decreased by filter mismatch.

Recall now that in addition to requiring $(x,x_*) = (y,y_*) = \alpha$ (for symmetry) we require $\alpha > \frac{1}{\sqrt{2}}$ to avoid having $P_e > \frac{1}{2}$. Since $||x_*|| = ||y_*|| = 1$, $\alpha^2 + \beta^2 < 1$, and since $\alpha > \frac{1}{\sqrt{2}}$, $|\beta| < \alpha$.

We seek the minimum of $\psi(\beta)$ over all β (i.e. no constraints) by differentiating ψ with respect to β, and find $\hat{\beta}$ such that $\left.\frac{d\psi}{d\beta}\right|_{\beta = \hat{\beta}} = 0$. If $\left.\frac{d^2\psi}{d\beta^2}\right|_{\beta = \hat{\beta}} > 0$, then $\hat{\beta}$ provides a minimum. We must now show that $\hat{\beta}$ is permissible, i.e. $(\alpha^2 + \hat{\beta}^2 < 1$ for $\alpha > \frac{1}{\sqrt{2}}$, and that the extreme values of β; $\beta_1 = \sqrt{1 - \alpha^2}$, $\beta_2 = \pm\sqrt{1 - \alpha^2}$ provide values of ψ that are greater than $\psi(\hat{\beta})$.

We find that

$$\hat{\beta} = \frac{1 - \alpha^2}{\alpha} \tag{2-66}$$

$$\psi(\hat{\beta}) = \sqrt{2\alpha^2} - 1 \tag{2-67}$$

$$\left.\frac{d^2\psi}{d\beta^2}\right|_{\beta = \hat{\beta}} > 0 \text{ for } \alpha^2 + \beta^2 < 1 \tag{2-68}$$

and indeed

$$\alpha^2 + \hat{\beta}^2 = \alpha^2 + \frac{1}{\alpha^2} - 2 + \alpha^2 \leq 1 \tag{2-69}$$

$$\text{or } 2\alpha^2 + \frac{1}{\alpha^2} \leq 3 \quad \text{for } \alpha \geq \frac{1}{\sqrt{2}} \tag{2-70}$$

also

$$\psi(\beta_1) = \psi(\beta_2) = 1 > \sqrt{2\alpha^2} - 1 \quad \text{for } \frac{1}{\sqrt{2}} < \alpha < 1 \tag{2-71}$$

so that $\hat{\beta}$ lies in the constraint set: Thus

$$\sqrt{2\alpha^2} - 1 = \sqrt{2\left[1 - \frac{\epsilon^2}{2}\right]^2} - 1 \leq \gamma \leq 1 \tag{2-72}$$

the left hand side of (2-71) represents the worst case degradation caused by filter mismatch error. It is easy to show that

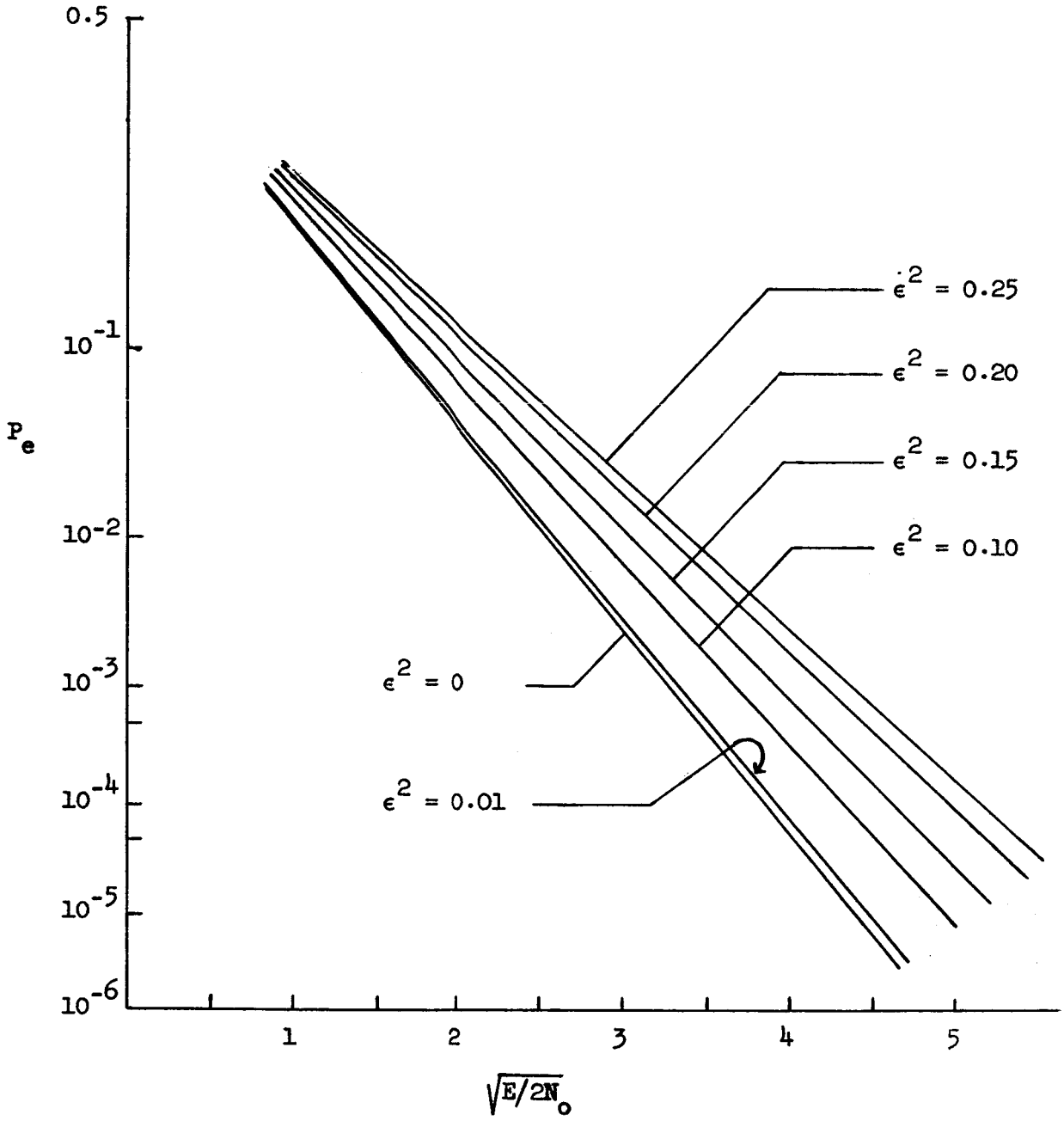


Fig. 2-6
Probability of Error for Mismatched Filters and Ideal Signals; Equal Energy Orthogonal Signals (Worst Case)

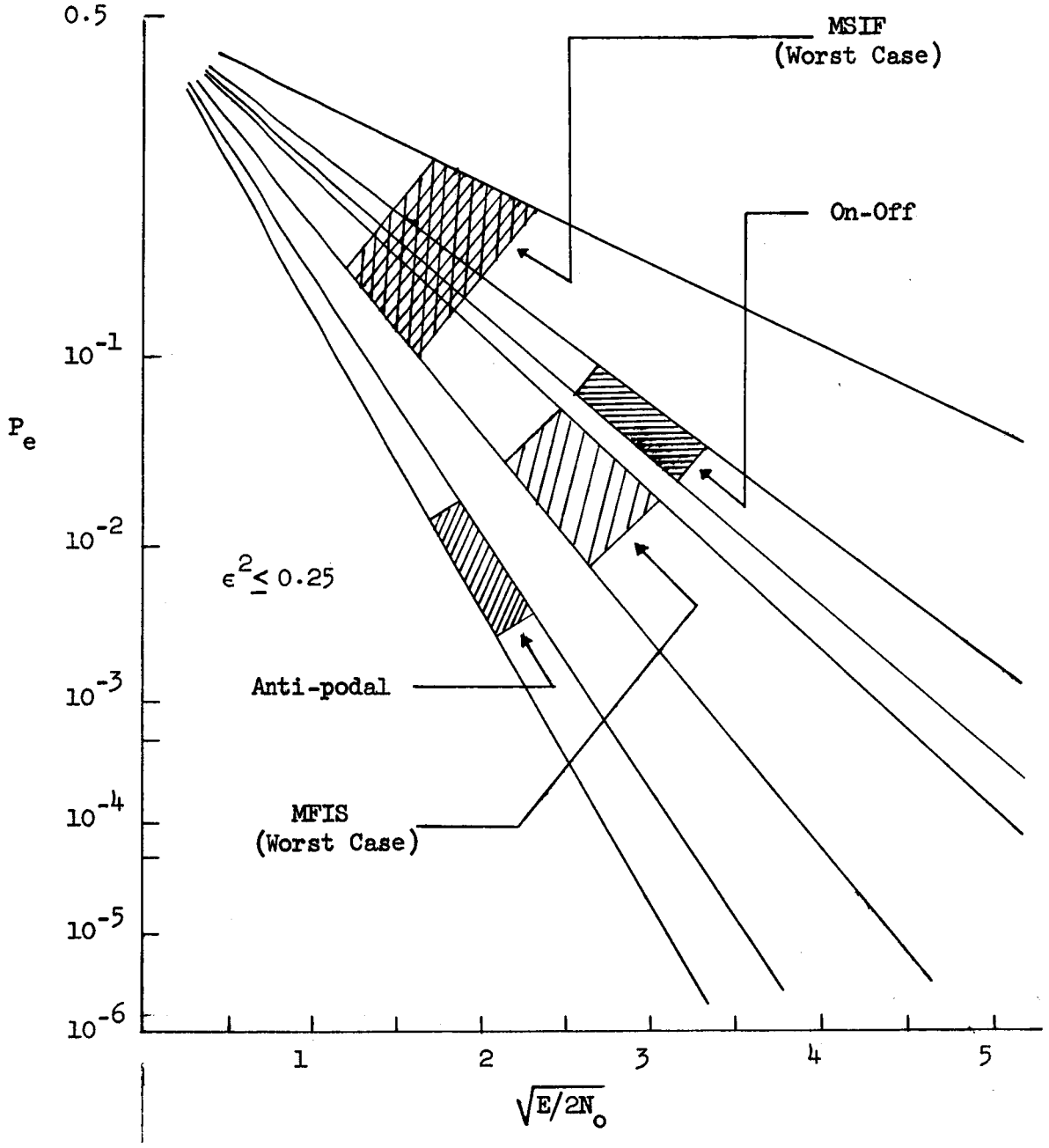


Fig. 2-7

Comparision of Degradation of Performance of Anti-podal, On-Off, and Orthogonal Signal Systems. ϵ^2.

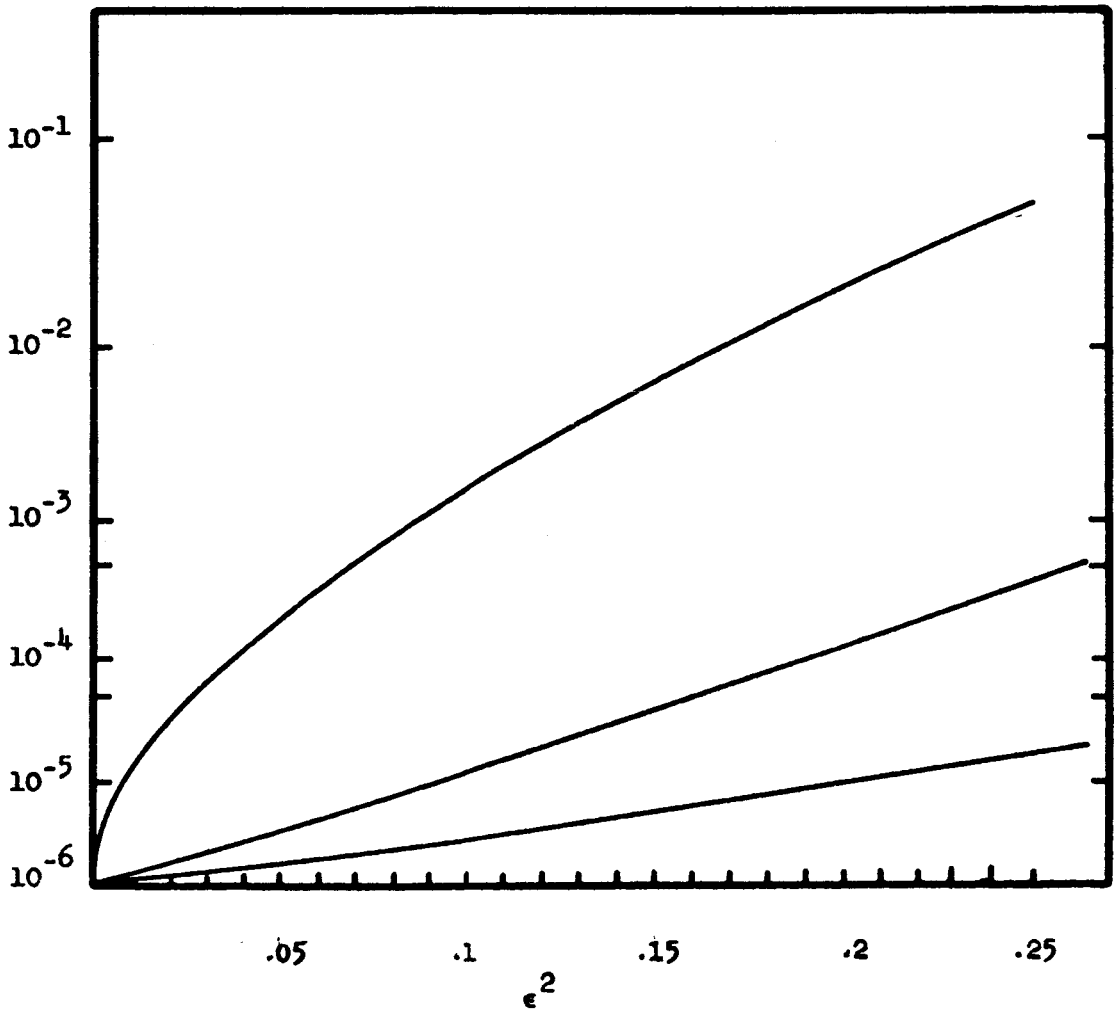


Fig. (2-8)

Rate of degredation of probability of error from an initial value of 10^{-6}

$$\gamma > \rho \qquad (2\text{-}73)$$

and thus the orthogonal signal system is less sensitive to filter mismatch error (for perfect signals) than to signal mismatch error (for perfect filters). This is indicated in Figures (2-6), (2-7) and (2-8).

2.9 Summary of Matched Filter Approximation

We have shown here the effect of signal and filter mismatch error on the performance (probability of error) of three matched filter receivers. For the "on-off" and anti-podal cases, it was found that only the magnitude (not the character) of the error influenced the performance. Moreover, for these two cases, it makes no difference whether the error is due to the fact that the actual received signal is not the same as the intended signal, or whether the filter is not quite matched to the received signal. The waveform of the approximate signal or filter is of no consequence. The equal energy orthogonal signal case is more interesting. Here the character or waveform of the error is important. Given only the magnitude of the mismatch error, one can only give upper and lower bounds on the performance. In contrast to the other two cases, where mismatch error always causes degradation in performance, a given mismatch error may leave performance unchanged. Also the location of the error (in the signal or in the filter) makes a considerable difference in the performance of the system. The decrease in performance when the filters are in error is less rapid than for the same error in the signal (comparing worst cases). One would infer from these results that of the three systems, the Orthogonal Signal System is potentially the most sensitive to mismatch error, and is potentially loss sensitive to filter mismatch them to signal mismatch. Signal mismatch error sometimes occurs due to waveform distortion caused by the channel, and in general, orthogonal transmitted signals do not result in orthogonal received signals. Some aspects of this problem are considered in the next chapter.

Chapter 3

WAVEFORM CONSTRUCTION

3.1 Introduction:

When communication signals are transmitted through a channel, the waveform of the output signal will, in general, differ from the waveform of the input signal. In particular, orthogonal input signals do not, in general, produce orthogonal output signals. Moreover, we cannot usually write an expression for the impulse response of the channel and determine analytically the effect of the channel on the transmitted signals. In this chapter we develop techniques that allow construction of a set of transmitted signals from an arbitrary set of basis functions, so that the set of output signals have a prescribed inner product matrix.* It is required only that the channel be linear and that the inner products of the output signals can be measured (by any convenient method such as time sampling.)

3.2 Waveform Construction

By "waveform construction" is meant here the formulation of a waveform by a linear combination of other waveforms. That is,

$$f(t) = \sum_{i=1}^{N} a_i \, g_i(t) \qquad (3\text{-}1)$$

---

* The inner product definition used for the output signals need not be the same as that used for the input signals.

where the g_i are arbitrary real functions and the a_i are real numbers. $f(t)$ is not being approximated by this linear combination; it is defined by it. The approximation problem is considered in the next chapter.

Here the waveform of $f(t)$ i.e. "what $f(t)$ looks like" is considered to be of little importance, and the g_i's may be selected on the basis of their case of generation, etc.

It will be found to be convenient if in eq. (3-1), the g_i's are orthonormal. How or where one finds a set of orthonormal signals provides a starting point for this discussion. One might pick a set of functions that is known to be orthogonal such as

$$\{g_n(t) = \sin \frac{n\pi}{T} t\} \qquad 0 \leq t \leq T \tag{3-2}$$

where as before, we define $(g_n, g_m) = \int_0^T g_n(t)\, g_n(t)\, dt$. That is, if $g(t)$ is given as

$$g(t) = \sin \frac{\pi}{T} t \tag{3-3}$$

and if one asks for some $h(t)$ such that $(h, g) = 0$, a "natural" choice would be another sine wave with frequency an integer multiple of $\frac{\pi}{T}$. If the given function is say

$$g(t) = |J_1(1.3t)|^{3/2} \operatorname{sgn}\left[\sum_{n=1}^{M} \frac{1}{n} \sin \frac{n\pi}{T} t\right] |jt(t-T)|^{1/2} \tag{3-4}$$

Then the choice of an $h(t)$ such that $(h, g) = 0$ becomes somewhat more difficult. Here if $h(t) = \frac{d}{dt} g(t)$, then $(h, g) = 0$ since for any integrable, differentiable function f such that $f(0) = f(T) = 0$,

$$\int_0^T f(t)\, f(t)\, dt = \frac{1}{2} f^2 \Big|_0^T = \frac{1}{2} [f(0) - f(T)] = 0 \tag{3-5}$$

If the given function is not recognized as belonging to some known set of orthogonal functions, or doesn't satisfy the consitions for some integration trick such as the above, the most straightforward way of constructing h so that (h, g) = 0 is the Gram-Schmidt orthonormalization process [8] which we now explain. It was remarked above that eq. (3-1) would be more convenient if the N f_i's were orthonormal. The Gram-Schmidt process is a method of forming N orthonormal functions from a set of N linear independent functions. A set of functions $\{f_1, f_2, \cdots f_N\}$ is said to be linearly independent if the relation $\sum_{i=1}^{N} c_i f_i(t) = 0$ for every t implies that $c_1 = c_2 = \cdots = c_N = 0$. Or in other words none of the N f_i's can be expressed as linear combination of the (N-1) other f_i's. The method is as follows: Given the linearly independent functions $f_1, \cdots f_N$, a set of orthonormal functions $\varphi_1, \varphi_2 \cdots \varphi_N$ is constructed, beginning by setting $\varphi_1 = \dfrac{f_1}{||f_1||}$. Clearly $||\varphi_1|| = 1$. Consider now the function $f_2' = f_2 - \lambda_1 \varphi_1$. Here λ_1 can be so chosen that the function f_2' is orthogonal to φ_1. Such is the case for $\lambda_1 = (f_2, \varphi_1)$. Consequently $f_2' = f_2 - (f_2, \varphi_1)\varphi_1$. If now we set

$$\varphi_2 = \frac{f_2'}{||f_2'||} \tag{3-6}$$

Then

$$||\varphi_2|| = 1, \tag{3-7}$$

$$\text{and} \quad (\varphi_2, \varphi_1) = 0$$

Next, form the function

$$f_3' = f_3 - (f_3, \varphi_1)\varphi_1 - (f_3, \varphi_2)\varphi_2 \tag{3-8}$$

which is orthogonal to the functions φ_1, φ_2. Now set

$$\varphi_3 = \frac{f_3'}{||f_3||} \tag{3-9}$$

and

$$||\varphi_3|| = 1, \ (\varphi_3, \varphi_1) = 0, \ (\varphi_3, \varphi_2) = 0 \tag{3-10}$$

In general if we put

$$f_k' = f_k - (f_k, \varphi_1)\varphi_1 - (f_k, \varphi_2)\varphi_2 - \cdots - (f_k, \varphi_{k-1})\varphi_{k-1} \tag{3-11}$$

and

$$\varphi_k = \frac{f_k'}{||f_k'||} \tag{3-12}$$

The set of functions $\varphi_1, \cdots \varphi_N$ is orthonormal. An explicit expression for the function φ_k is given by

(3-13)

$$\varphi_k = \frac{1}{\sqrt{F_k F_{k-1}}} \begin{vmatrix} (f_1,f_1) & (f_2,f_1) & \cdots & (f_k,f_1) \\ (f_1,f_2) & (f_2,f_2) & \cdots & (f_k,f_2) \\ (f_1,f_{k-1}) & (f_2,f_{k-1}) & \cdots & (f_k,f_{k-1}) \\ f_1 & f_2 & \cdots & f_k \end{vmatrix}$$

$$F_o = 1 \quad F_k = F(f_1, \cdots f_k) \quad K = 1, 2, \cdots N \tag{3-14}$$

where

$$F_k = \begin{vmatrix} (f_1,f_1) & (f_2,f_1) & \cdots & (f_k,f_1) \\ (f_1,f_2) & (f_2,f_2) & \cdots & (f_k,f_2) \\ \cdots & \cdots & \cdots & \cdots \\ (f_1,f_k) & (f_2,f_k) & \cdots & (f_k,f_k) \end{vmatrix} \tag{3-15}$$

and is called Grams determinant of the set of functions $f_1, \cdots f_k$.

The relationship between the φ's and f's is seen to be of the form

$$\begin{aligned} \varphi_1 &= a_{11} f_1 \\ \varphi_2 &= a_{21} f_1 + a_{22} f_2 \\ &\cdots \\ \varphi_N &= a_{N1} f_1 + a_{N2} + \cdots + a_{NN} f_N \end{aligned} \tag{3-16}$$

or in matrix notation,

$$\Phi = A F \tag{3-17}$$

where

$$\Phi = \begin{bmatrix} \varphi_1 \\ \cdot \\ \cdot \\ \cdot \\ \varphi_N \end{bmatrix} \tag{3-18}$$

$$A = \begin{bmatrix} a_{11} & 0 & 0 \cdots & 0 \\ a_{21} & a_{22} & 0 \cdots & 0 \\ \cdot & \cdot & \cdot \cdot \cdot & \cdot \\ a_{N1} & a_{N2} & \cdots\cdots & a_{NN} \end{bmatrix} \tag{3-19}$$

and

$$F = \begin{bmatrix} f_1 \\ \cdot \\ \cdot \\ \cdot \\ f_N \end{bmatrix} \tag{3-20}$$

It is seen then that given any integrable function on [0, T] there are infinitely many functions that are orthogonal to f_1, as there are infinitely many functions which may be selected as f_2 in the Gram-Schmidt process.

Given a set of N functions, the labeling (calling one of the functions f_1, another f_2, etc.) may be done in N! different ways since there are N choices for f_1, N-1 choices for f_2 having chosen f_1, etc. This means that there are N! different Φ's that may be constructed from a set of N linearly independent functions, although they span the same space.

3.3 Signals for M-ary Systems

If two known signals, x_1 and x_2, are received in the presence of white gaussian noise, it is well known that the signal selection problem reduces to choosing $x_1 = -x_2$ in order to minimize probability of error. If the receiver must decide which one of M possible signals was sent, it has been conjectured [9], but only recently proved [10], that one should choose a set of signals having the largest negative pairwise correlation (or inner-product). For a set of N signals having unit norm (or energy), the largest negative pairwise correlation is given by

$$\rho = \frac{-1}{N-1} \cdot \tag{3-21}$$

This is shown simply by noting that if $f_1, \ldots f_N$ are signals having unit norm and $(f_i, f_j) = \rho$ $i \neq j$ then

$$\left(\sum_{i=1}^{N} f_i, \sum_{i=1}^{N} f_i \right) \geq 0 \tag{3-22}$$

$$= \sum_{i}^{N} \sum_{j}^{N} (f_i, f_j) \tag{3-23}$$

$$= N + (N^2-N)\,\rho \geq 0$$

$$\text{or } \rho \geq \frac{-1}{N-1} \cdot \tag{3-24}$$

Equality holds here if and only if $\sum_{i=1}^{N} f_i = 0$. Note that equality here implies that $f_1, \ldots, f_N$ are linearly dependent.

We now show that $\sum_{i=1}^{N} f_i$ is the only linear combination of the f_i that vanishes if the f_i have unit norm and $\rho = (f_i, f_j) = \frac{-1}{N-1}$. Suppose that $\sum_{i=1}^{N} \lambda_i f_i(t) = 0$ and $\rho = \frac{-1}{N-1}$. Then for any $f_k, 1 \leq k \leq N$, $(f_k, \sum_{i=1}^{N} \lambda_i f_i) = 0$ since $\sum_{i=1}^{N} \lambda_i f_i = 0$. (3-25)

$$\text{But } (f_k, \sum_{i=1}^{N} \lambda_i f_i) = \sum_{i=1}^{N} \lambda_i (f_i, f_k) = \lambda_k \frac{-1}{N-1} \sum_{i \neq k}^{N} \lambda_i \tag{3-26}$$

$$= \lambda_k \left[1 + \frac{1}{N-1}\right] - \frac{1}{N-1} \sum_{i=1}^{N} \lambda_i \tag{3-27}$$

$$= \frac{N}{N-1} \left[\lambda_k - \frac{1}{N} \sum_{i=1}^{N} \lambda_i\right]. \tag{3-28}$$

Since this is zero for all k, all the λ'_ks are equal, so that the only linear combination of the f_i to vanish is $\sum_{i=1}^{N} f_i$ (or a scalor multiple thereof). In other words, if N-1 signals $f_1, \ldots, f_{N-1}$ of unit norm are constructed having $(f_i, f_j) = \frac{-1}{N-1}$, the only f_N having the property that $(f_N, f_i) = \frac{-1}{N-1}$ i=1, ...,N-1 is

$$f_N = -\sum_{i=1}^{N-1} f_i. \tag{3-29}$$

That $||f_N|| = 1$ is shown simply by writing $||f_N||^2 = ||\sum_{i=1}^{N-1} f_i||^2 =$

$$\sum_{i=1}^{N-1} ||f_i||^2 + \sum_{i \neq j}^{N-1} \sum^{N-1} (f_i, f_j \tag{3-30}$$

and using the fact that $||f_i|| = 1$ and $(f_i, f_j) = -\frac{1}{N-1}$,

we have $||f_N||^2 = (N-1)\,1 + \{[N-1]^2 - [N-1]\}\left[\frac{-1}{N-1}\right] = 1$ (3-31)

The fact that the negative of the sum of N-1 signals of unit norm having $\rho = \frac{-1}{N-1}$ is the only signal f_N having the property that $(f_N, f_i) = \frac{-1}{N-1}$ will be made use of in a construction procedure developed later in this chapter.

It has been correctly pointed out, [12], that equality in eq. (3-23) can be achieved if and only the number of signals (N) is at least one greater than the dimensionality of the signal space. If N f_i are considered to be constructed from M orthonormal signals φ_i:

$$f_i = \sum_{i=1}^{M} a_{1i}\,\varphi_i$$

. . . . . .

$$f_N = \sum_{i=1}^{M} a_{Ni}\,\varphi_i .$$

The dimensionality of the signal space is M (if no $a_{ki} = 0$ for every k).

A technique of constructing a set of signals having this minimum equal correlation property (called Regular-simplex Codes) is described by Stutt [13]. Another method is presented at the end of this chapter. In [13], the signals are considered to be vectors whose components are time samples of the continuous waveforms. This is an unnecessary restriction, and his method applies equally well to the more general formulation of (eq. 3-24). If time samples are used for the coordinates then the φ_i may be taken as non-overlapping pulses, producing f_i's which are staircase functions or the f_i may be assumed to be bandlimited functions and the φ_i are sinc functions ($\varphi = \frac{\sin x}{x}$). In general; if the f_i are known to have the form of eq. 3-24, a_{ki} is given by (f_k, φ_i). For the above two special cases, (f_k, φ_i) is proportional to the time samples of the f's.

3.4 Linear Filtering of Constructed Signals

Unless a set of signals with a prescribed correlation or inner product matrix is available to the signal designer, it is his task to build the desired set. The construction of an orthogonal set of signals from a linearly independent set has been illustrated previously with the Gram-Schmidt Orthnormalization Process. One construction procedure for generating a so called regular simplex code, as metioned before, has been given by Stutt; we give another in this chapter. As was noted before, restriction to time sampling is not necessary, and Stutt's construction procedure holds for the more general formulation given earlier, where we write

$$\begin{aligned} f_1 &= \sum^{N} a_{1i}\, \varphi_i \\ &\vdots \\ f_M &= \sum^{M} a_{Mi}\, \varphi_i \end{aligned} \qquad (3\text{-}32)$$

and the φ_i's are arbitrary orthonormal functions.

If these signals are to be used in the usual model of a communication channel, where the transmitted and the received waveforms are the same, then the choice of the φ_i's is immaterial. The performance of the system depends only on the inner product matrix of the set of signals, and as is shown by Balak [10], the probability of error is minimized if the set of signals is a regular simplex code. For N fairly large (say 10 or more), orthogonal signals perform about as well and are apparently considerably easier to generate.

In eq. (3-32) the φ_i are also constructed signals unless they are given as being orthonormal. In other words the φ_i's are constructed from a

linearly independent set of functions $g_1, g_2 \cdots g_k$. In matrix notation, we write eq. (3-32) as

$$f = A\,\Phi \tag{3-33}$$

Then writing

$$\Phi = B\,G \tag{3-34}$$

$$F = A\,B\,G \tag{3-35}$$

where

$$F = [f_1, f_2 \cdots f_M]^T \tag{3-36}$$

$$\Phi = [\varphi_1, \varphi_2 \cdots \varphi_N]^T \tag{3-37}$$

$$G = [g_1, g_2 \cdots g_k]^T \tag{3-38}$$

$$K \geq N$$

A is an N x M matrix, and B is an N x N matrix. In eq. (3-32) N has been called the dimensionality of the signals f_i. A detailed discussion of dimensionality is taken up in the next chapter; it suffices here to note that one must speak of the dimensionality of a set of functions, not the dimensionality of a single function. A single function makes up a one point set and is one dimensional, regardless of how it may be decomposed. That is, the fact that a function f may be written as

$$f(t) = \sum_{1}^{N} f\left(\frac{n}{2\omega}\right) \frac{\sin\left(t - \frac{n}{2\omega}\right)}{t - \frac{n}{2\omega}} \tag{3-39}$$

does not make f and N dimensional signal. The dimensionality, (as defined in Chapter V) of the set of f's in eq. (3-32), would be the smaller of M and N.

As was stated at the beginning of this section, no account is usually taken of the distortion of the signal waveshape that may take place during the passage from transmitter to receiver. Decomposition of a set of signals as in eq. (3-32) may be done for purposes of construction, or for ease of

finding the response of the channel to the constructed waveforms.

If two transmitted signals are designed to be orthonormal, the received signals in general no longer have this property. Let the linear channel be represented by its impulse response h(t). If φ_1 and φ_2 are two orthonormal transmitted signals, the received signals are given by

$$\theta_1(t) = \int_T h(t-\tau)\, \varphi_1(\tau)\, d\tau \tag{3-40}$$

$$\theta_2(t) = \int_T h(t-\tau)\, \varphi_2(\tau)\, d\tau \tag{3-41}$$

Then $(\theta_1, \theta_2) = \int_T \theta_1(t)\, \theta_2(t)\, dt$

$$= \int_T \int_T \int_T h(t-\tau)\, h(t-\beta)\, \varphi_1(\tau)\, \varphi_2(\beta)\, d\tau\, d\beta$$

Following the notation of Chap. II,

$$(\theta_1, \theta_2) = \int_T \int_T H(\tau,\beta)\, \varphi_1(\tau)\, \varphi_2(\beta)\, d\tau\, d\beta \tag{3-42}$$

Now if φ_1, φ_2 are selected to be eigenfunctions corresponding to distinct λ's of the integral equation

$$\lambda\, \varphi(\tau) = \int_T H(\tau,\beta)\, \varphi(\beta)\, d\beta \tag{3-43}$$

we have

$$\int_T \left\{ \int_T H(\tau,\beta)\, \varphi_1(\tau)\, d\tau \right\} \varphi_2(\beta)\, d\beta$$

$$= \int_T \lambda_1\, \varphi_1(\beta)\, \varphi_2(\beta)\, d\beta = 0 \tag{3-44}$$

since the eigenfunctions of eq. (3-43) are orthogonal. That is, a sufficient condition for a set of received signals to be orthogonal when their corresponding transmitted signals are orthogonal, is that the transmitted signals satisfy

eq. (3-43). That it is not necessary is seen by considering the transmitted signals to be $\varphi_1 = x$, $\varphi_2 = \dot{x}$, with $x(0) = x(T)$. Then $(\varphi_1, \varphi_2) = 0$, and since the channel is linear $h[x] = y$ implies $h[\dot{x}] = \dot{y}$, so that if $y(0) = y(T) = 0$, $(y, \dot{y}) = 0$. Another, less artificial, example appears later in the chapter.

3.5 Construction of Transmitted Signals

Here the transmitted (input) signals are passed through a linear channel h. If the input signals have a certain inner product matrix (e.g. a regular simplex-code), the output (received) signals will not in general have the same inner product matrix. It is of interest to determine the relationship between these two matrices for a given channel. In particular, we give method for constructing a set of input signals so that the set of output signals has a prescribed inner product matrix. As a special case of this method, a very simple procedure for constructing regular simplex codes is presented.

Let the set of constructed input signals be given by

$$\begin{aligned} f_1 &= \sum_{i=1}^{N} a_{1i}\, \varphi_i \\ &\vdots \\ f_M &= \sum_{i=1}^{N} a_{Mi}\, \varphi_i \end{aligned} \qquad (3\text{-}45)$$

where the φ_i's are arbitrary orthonormal signals. Call θ_i the response of the linear channel h to φ_i, and g_i the response of h to f_i. The set of output signals is then given by

$$\begin{aligned} g_1 &= \sum_{i=1}^{N} a_{1i}\ \theta_i \\ &\vdots \\ g_M &= \sum_{i=1}^{N} a_{Mi}\ \theta_i \end{aligned} \tag{3-46}$$

Let

$$b_{ij} = (\theta_i,\ \theta_j) \tag{3-47}$$

$$g_{k\ell} = (g_k, g_\ell) = \sum_{i=1}^{N}\sum_{i=1}^{N} a_{ki}\ a_{\ell j}\ b_{ij} \tag{3-48}$$

We introduce the notation $< GG^T > = G^* = <$

$$\begin{bmatrix} g_1\ g_1 & \cdots & g_1\ g_M \\ g_M\ f_1 & \cdots & g_M\ g_M \end{bmatrix} > \tag{3-49}$$

$$= \begin{bmatrix} (g_1,\ g_1) & \cdots & (g_1,\ g_M) \\ (g_M,\ g_1) & \cdots & (g_M,\ g_M) \end{bmatrix}$$

The inner product matrix of the q's may be written as

$$< GG^T > = G_* = [(g_k,\ g_\ell)] = AB_*A^T \tag{3-50}$$

where

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & & \\ a_{M1} & \cdots & a_{MN} \end{bmatrix} \tag{3-51}$$

is the generator matrix for the f's ($F = A\Phi$) and

$$B_* = \begin{bmatrix} (\theta_1,\ \theta_1) & \cdots & (\theta_1,\ \theta_N) \\ \vdots & & \\ (\theta_N,\ \theta_1) & \cdots & (\theta_N,\ \theta_N) \end{bmatrix} \tag{3-52}$$

is the inner product matrix of the θ's.

The inner product matrix of the set of input signals is given by

$$F_* = AA^T \tag{3-53}$$

and the condition for $G_* = KF_*$, i.e., the inner product matrix of the output signals is proportional to the inner product matrix of the input signals is that

$$AB_*A^T = KAA^T = F_* \tag{3-54}$$

$$\text{or } B_* = KI \tag{3-55}$$

where I is the identity matrix, and K is an arbitrary constant. If the channel is anything but an attenuator, then the two matirces are not the same, and it is of interest to determine a method for constructing the input signals so that the output signals have a prescribed inner product matrix.

We make use of the fact (theorem 1, p.126 [32]) that a non-singular symmetric matrix may be uniquely deocmposed into the product of a lower triangular matrix and its transpose. That is, if X is an (NXN) symmetric matrix, it may be written as

$$X = TT^T \tag{3-56}$$

where

$$T = \begin{bmatrix} t_{11} & 0 & \cdots & 0 \\ t_{21} & t_{22} & \cdots & \vdots \\ \vdots & & & \vdots \\ t_{N1} & t_{N2} & \cdots & t_{NN} \end{bmatrix} \tag{3-57}$$

and T is unique. We require here that A be square and non-singular.

Let

$$G_* = \alpha\alpha^T \text{ and } \beta_* = \beta\beta^T \tag{3-58}$$

where α and β are lower triangular matrices. Then Eq. (3-50) becomes

$$\alpha\alpha^T = A\beta\beta^T A^T \tag{3-59}$$

$$\text{or } \alpha\alpha^T = (A\beta)(A\beta)^T \tag{3-60}$$

We now identify α and $A\beta$. That is, we set

$$\alpha = A\beta \tag{3-61}$$

$$\text{or } A = \alpha\beta^{-1} \tag{3-62}$$

Equation (3-62) assumes a particlarly simple form if the φ's are selected according to Eq. (3-43). As was noted previously, the θ's are then orthogonal, with

$$(\theta_i, \theta_j) = \lambda_i \delta_{ij} \tag{3-63}$$

and the inner product matrix of the θ's is given by

$$B_* = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_N \end{bmatrix} \tag{3-64}$$

Then

$$B = \begin{bmatrix} \sqrt{\lambda_1} & & & 0 \\ & \sqrt{\lambda_2} & & \\ & & \ddots & \\ & & & \sqrt{\lambda_N} \end{bmatrix} \tag{3-65}$$

The inverse of β is given by

$$\beta^{-1} = \begin{bmatrix} \frac{1}{\sqrt{\lambda_1}} & & & \\ & \frac{1}{\sqrt{\lambda_2}} & & 0 \\ & & \ddots & \\ 0 & & & \frac{1}{\sqrt{\lambda_N}} \end{bmatrix} \tag{3-66}$$

A procedure for finding the triangular decomposition of a symmetric matrix is given on page 127 of [32]. The elements of the lower triangular matrix is calculated as follows: If the symmetric matrix G has elements $g_{ik} = g_{ki}$, the elements α_{ik} ($\alpha_{ik} = 0$ for $k > i$) of the lower triangular matrix are computed in the following fashion.

First column: $\alpha_{11} = \sqrt{g_{11}}$

$$\alpha_{k1} = \frac{g_{k1}}{\alpha_{11}}$$

2^{nd} column : $\alpha_{22} = \sqrt{(g_{22} - \alpha_{21}^2)}$

$$\alpha_{k2} = (g_{k2} - \alpha_{21}\,\alpha_{k1})\,/\alpha_{22}$$

3^{rd} column : $\alpha_{33} = \sqrt{g_{33} - \alpha_{31}^2 - \alpha_{32}^2}$ (3-67)

$$\alpha_{k3} = (g_{k3} - \alpha_{31}\,\alpha_{k1} - \alpha_{32}\,\alpha_{k2})/\alpha_{33}$$

4^{th} column : $\alpha_{44} = \sqrt{g_{44} - \alpha_{41}^2 - \alpha_{42}^2 - \alpha_{43}^2}$

$$\alpha_{k4} = (g_{k4} - \alpha_{41}\,\alpha_{k1} - \alpha_{42}\,\alpha_{k2} - \alpha_{43}\,\alpha_{k3})/\alpha_{44}$$

etc.

3.6 A Construction Procedure for Regular Simplex Codes

Here it is desired to construct a set of N functions $\{f_i, \cdots f_N\}$ so that

$$(f_i, f_j) = \begin{cases} 1 & i = j \\ \frac{-1}{N-1} & i \neq j \end{cases} \tag{3-68}$$

If the f's are generated by a set of orthonormal functions,

$$F = A\Phi \tag{3-69}$$

Then as before, the inner product matrix of the f's is given by

$$FF^T = F_* = [(f_i, f_j)] = AA^T \tag{3-70}$$

Here we simply identify A with the lower left triangular matrix of the triangular decomposition of $[(f_i, f_j)]$ according to the procedure in Eq. (3-67).

However, in this case, the matrix F_* is singular since its determinant must vanish as the f's are linearly dependent. To avoid any possible formal difficulty in dealing with a singular matrix, and also to simplify the construction procedure we use the non-singular [N-1] by [N-1] matrix whose off-diagonal elements are equal to $\frac{-1}{N-1}$. This matrix is then factored into $A' A'^T$ where A' is a [N-1] by [N-1] lower triangular matrix. The generator matrix A for which AA^T yields the desired regular simplex code is found by adjoining to A' a row whose elements are equal to the negative of the sum of the elements in the respective columns. That this is correct is assured by the fact noted earlier that if $f_1, \cdots f_{N-1}$ have unit norm and $(f_i, f_j) = \frac{-1}{N-1}$, the only function f_N having the property that $(f_N, f_i) = \frac{-1}{N-1}$ is

$$f_N = - \sum_{i=1}^{N-1} f_i .$$

The above factorization procedure is of course not the only one that can be used, but it leads to perhaps the simplest computational procedure. This is especially true when the transmitted signals pass through a channel which changes their waveshape.

We now give some illustrative examples of the foregoing material. The above techniques do not require knowledge of the impulse response (or some equivalent characterization) of the channel. We only require that the channel's response to the set of input signals can be measured. In the first example, the channel is taken to be an RC lowpass circuit to permit analytic computation.

Example 4-1: Construction of input signal set to produce an orthonormal set of output signals:

Specifically in this example, we construct two input signals, f_1 and f_2, so that the output signals, g_1 and g_2 are orthonormal. Here the channel is taken to be an RC circuit with transfer function $H(s) = \frac{4}{s+4}$. The waveform of the φ's is arbitrary, but we take

$$\begin{aligned} \varphi_1 &= \sqrt{2}\, \epsilon^{-t} \\ \varphi_2 &= 2[3\epsilon^{-2t} - 2\epsilon^{-t}] \end{aligned} \tag{3-71}$$

for computational convenience. Then θ_1 and θ_2 (the channel's response to φ_1 and φ_2 respectively, are given by

$$\begin{aligned} \theta_1 &= \frac{4\sqrt{2}}{3}\ [\epsilon^{-t} - \epsilon^{-4t}] \\ \theta_2 &= \frac{4}{3}\ [-4\epsilon^{-t} + 9\epsilon^{-2t} - 5\epsilon^{-4t}] \end{aligned} \tag{3-72}$$

Here the inner product, x and y, (x,y) is taken to be

$$(x,y) = \int_0^\infty x(t)\, y(t)\, dt \tag{3-73}$$

We then find that

$$B_* = (\theta_i,\ \theta_j) = \begin{bmatrix} \frac{4}{5} & -\frac{2\sqrt{2}}{15} \\ -\frac{2\sqrt{2}}{15} & \frac{2}{3} \end{bmatrix} \tag{3-74}$$

and

$$\beta = \begin{bmatrix} \frac{4}{5} & & 0 \\ & & \\ \frac{-\sqrt{10}}{15} & \frac{2}{3} & \frac{1}{5} \end{bmatrix} \tag{3-75}$$

Since we want $(g_i, g_j) = \delta_{ij}$.

$$G_* = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{3-76}$$

and

$$\alpha = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{3-77}$$

so that

$$A = \alpha\beta^{-1} = \beta^{-1} \tag{3-78}$$

and

$$\beta^{-1} = \begin{bmatrix} \sqrt{\frac{5}{2}} & & 0 \\ & & \\ \frac{1}{4}\sqrt{\frac{10}{7}} & & \frac{3}{2}\sqrt{\frac{5}{7}} \end{bmatrix} = A \tag{3-79}$$

The operation is indicated schematically and pictorially in Figures (3-1) and (3-2).

Example 3-2. Construction of input signal set to produce a set of input signals which is a regular simplex code.

We take the same channel, and the same set of φ's as in Example (3-1), and construct three input signals so that the three output signals have the inner product matrix.

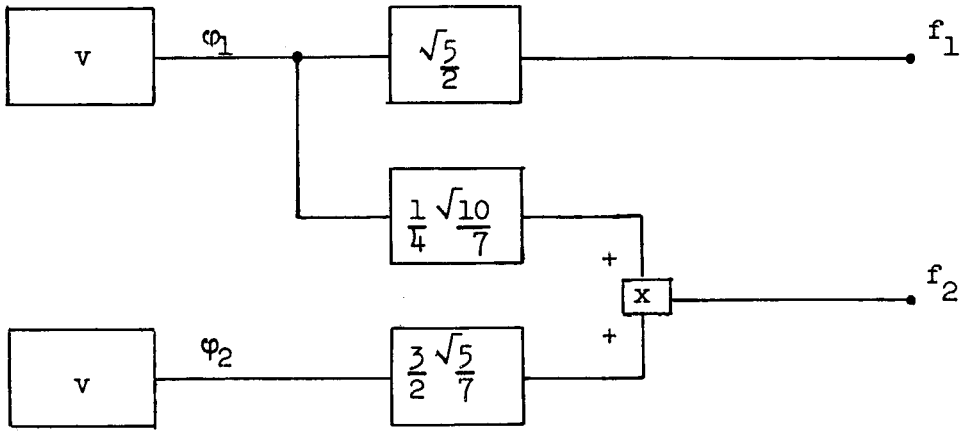


Fig. 3-1

Block Diagram of the Transmitter in Example

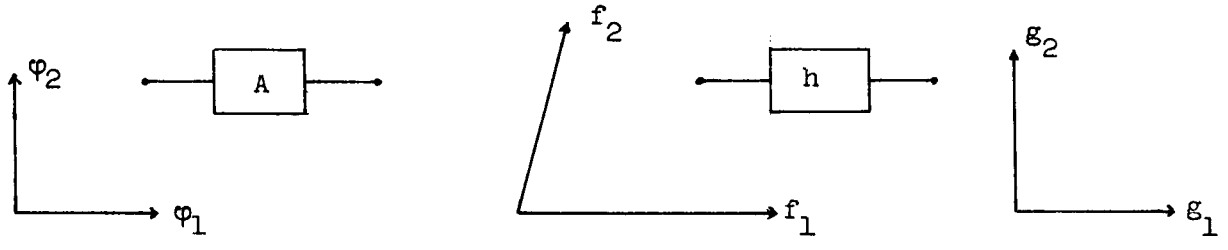


Fig. 3-2

Picorial Diagram of the Relations Between
the Signals in Examples (3-1)

$$G_* = [(g_i, g_j)] = \begin{bmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & 1 \end{bmatrix} \tag{3-80}$$

We make use of a previous result (eq. 3-28), that if a set of N-1 normal functions has the property

$$(f_i, f_j) = -\frac{1}{N-1} \quad i \neq j \; i, j = 1, ; -N\text{-}1, \tag{3-81}$$

the function f_N having the property that

$$(f_N, f_i) = -\frac{1}{N-1} \tag{3-82}$$

is given by $f_N = -\sum_{i=1}^{N-1} f_i$. We then construct two input signals, f_1 and f_2, so that their corresponding output signals, g_1 and g_2, have the property $(g_1, g_2) = -\frac{1}{2}$. By the above, $g_3 = -(g_1 + g_2)$, and since the channel is linear, $f_3 = -(f_1 + f_2)$. We consider then

$$G_* = \begin{bmatrix} 1 & -\frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix} \tag{3-83}$$

for which

$$\alpha' = \begin{bmatrix} 1 & 0 \\ \frac{1}{2} & \sqrt{\frac{3}{4}} \end{bmatrix} \tag{3-84}$$

then $A' = \alpha' \beta^{-1}$

$$= \begin{bmatrix} 1 & 0 \\ -\frac{1}{2} & \sqrt{\frac{3}{4}} \end{bmatrix} \begin{bmatrix} \sqrt{\frac{5}{2}} & 0 \\ \frac{1}{4}\sqrt{\frac{10}{7}} & \frac{3}{2}\sqrt{\frac{5}{7}} \end{bmatrix} \tag{3-85}$$

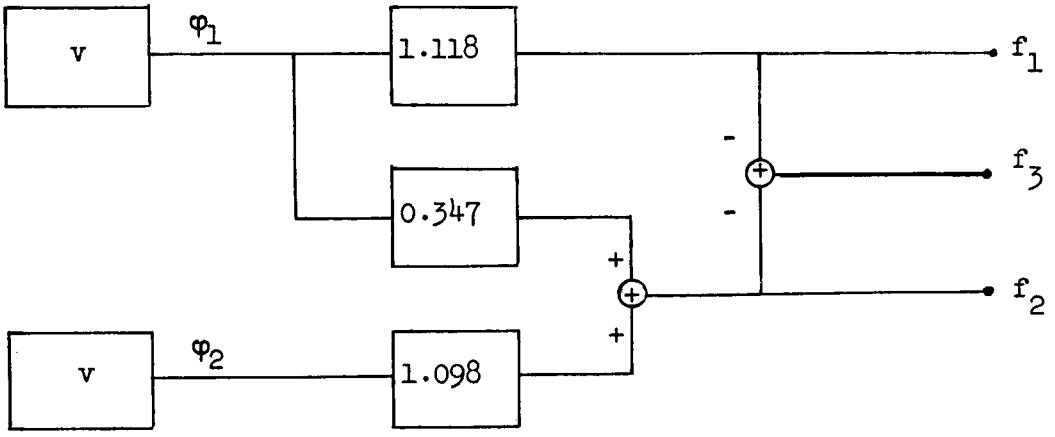


Fig. 3-3

Block Diagram of the Transmitter in Example (3-3)

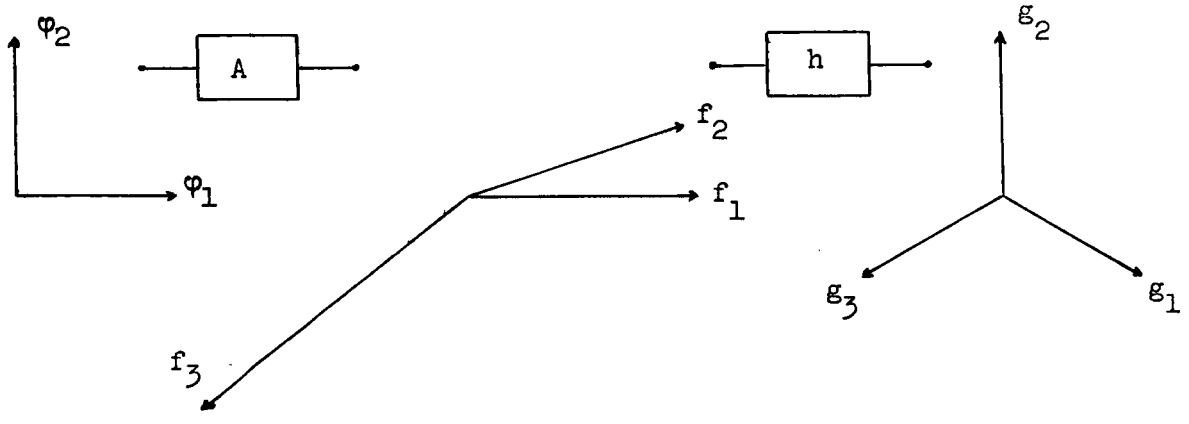


Fig. 3-4

Pictorial Diagram of the Relations between the Signals in Example (3-2)

$$= \begin{bmatrix} 1.118 & 0 \\ 0.347 & 1.098 \end{bmatrix} \tag{3-86}$$

A is found by adjoining to A′ a third row whose elements are given by the negative of the sum of elements of the respective columns. That is

$$A = \begin{bmatrix} 1.118 & 0 \\ 0.347 & 1.098 \\ -1.465 & -1.098 \end{bmatrix} \tag{3-87}$$

The operation is indicated schematically and pictorially in Figures (3-3) and (3-4).

Example 3-3

In the procedure given for construction a set of input signals so that the resulting output signals have a prescribed inner product matrix, it is not necessary that the inner product used for the input signals be the same as that for the output signals. That is, we may use for the φ's and f's the inner product

$$(x,y)_1 = \int_0^{T_1} x(t)\, y(t)\, dt \tag{3-88}$$

and for the θ's and g's the inner product

$$(x,y)_2 = \int_0^{T_2} x(t)\, y(t)\, dt \tag{3-89}$$

As an example we take φ_1 and φ_2 to be as shown in Fig. , and the channel to be an RC circuit with impulse response $h(t) = \epsilon^{-t}$. For the input signals, we take

$$(x,y)_1 = \int_0^{1} x(t)\, y(t)\, dt \tag{3-90}$$

and for the output signals,

$$(x,y)_2 = \int_0^\infty x(t)\, y(t)\, dt \tag{3-91}$$

the functions θ_1 and θ_2 are shown in Fig. (3-53). The inner product matrix of the θ's using $(\quad)_2$ is found to be

$$B = \begin{bmatrix} 0.36788 & 0 \\ 0 & 0.28382 \end{bmatrix} \tag{3-92}$$

Here the θ's turn out to be orthogonal, so that β and β^{-1} are diagonal.

$$\beta = \begin{bmatrix} 1.6487 & 0 \\ 0 & 1.8771 \end{bmatrix} \tag{3-93}$$

$$\beta^{-1} = \begin{bmatrix} 1.6487 & 0 \\ 0 & 1.8771 \end{bmatrix} \tag{3-94}$$

If we desire $(g_i, g_j) = \delta_{ij}$, then as before

$$A = \beta^{-1} \tag{3-95}$$

Since the θ's are orthogonal, this operation amounts to amplitude scaling of the input signals so that the output signals are normal.

If we require the output signals to be orthonormal on $0 \leq t \leq 1$, we compute $B = [\theta_i, \theta_j)_1]$:

$$B = \begin{bmatrix} 0.1681 & 0.0489 \\ 0.0489 & 0.04625 \end{bmatrix} \tag{3-96}$$

and

$$A = \beta^{-1} = \begin{bmatrix} 2.439 & 0 \\ -1.623 & 5.589 \end{bmatrix} \tag{3-97}$$

The input signals f_1 and f_2 are shown in Fig. 3-5.

The technique given above for finding a generator matrix A yields a triangular matrix. Once this triangular matrix is found, many different decomposition may, of course, be found by multiplying A by an orthogonal matrix.

Earlier in this chapter it was mentioned that one may choose to construct a set of input signals from another set because of the ease of physically generating the basic signals. One can, of course, only obtain signals that are linear combinations of the basic signals. For example, the signals which produces maximum energy transfer (as in Chapter III) may not lie in this class. We may look for an optimum signal in the class of signals generated by the baisic functions $\varphi_1, \cdots \varphi_N$, i.e.

$$f = \sum_{i=1}^{N} a_i \varphi_i \, . \tag{3-98}$$

In matrix notation we write

$$f = A\Phi \tag{3-99}$$

where $A = [a_1 \; a_2 \cdots a_N]$ and

$$\Phi = \begin{bmatrix} \varphi_1 \\ \cdot \\ \cdot \\ \varphi_N \end{bmatrix} \tag{3-100}$$

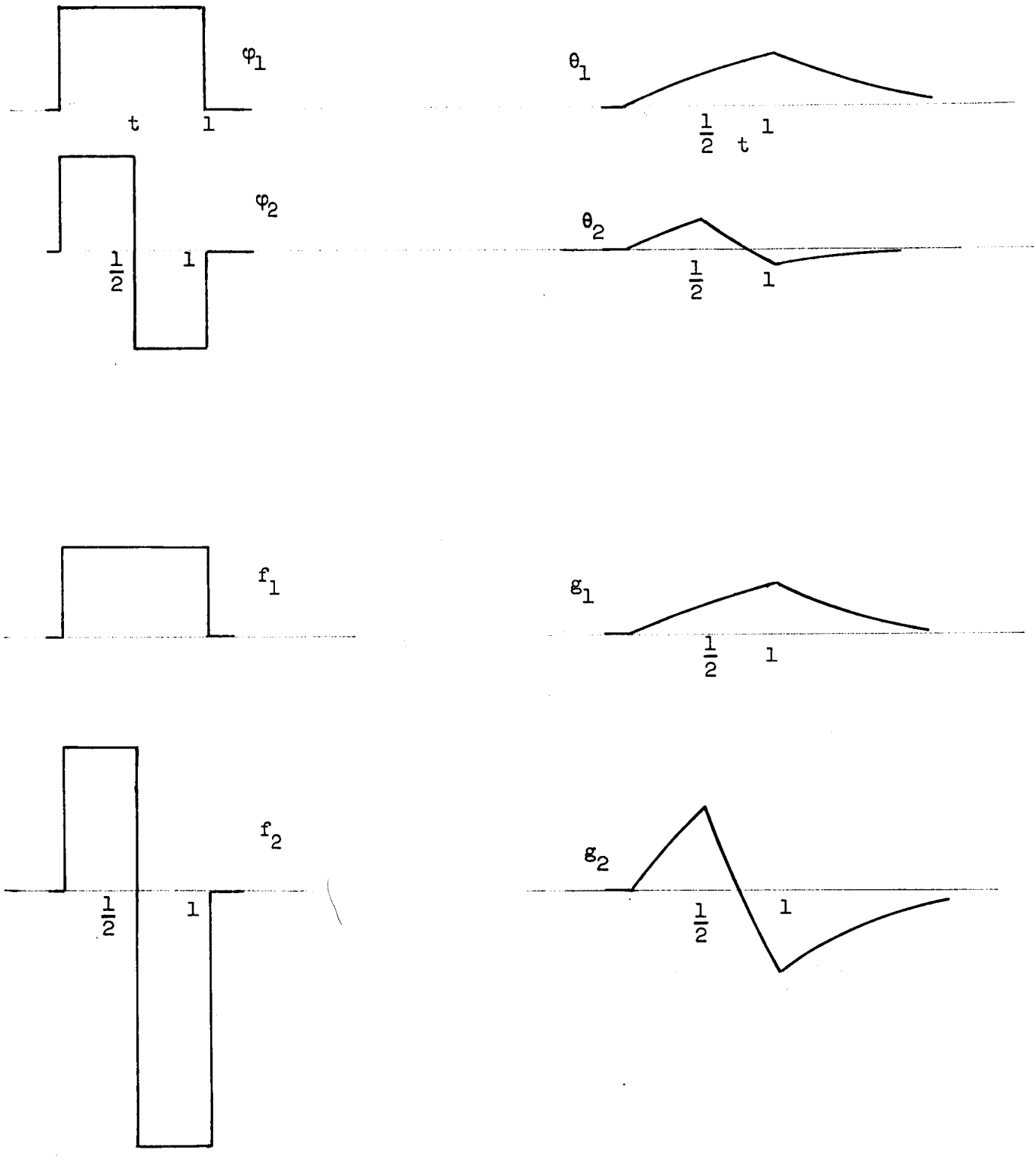


Fig. 3-5

Input and Output Waveforms of the Signals in Example (3-2)

We wish to maximize

$$\int_0^{T_2} g^2(t)\,dt = \int_0^{T_2} \Big(\sum_{i=1}^{N} a_i\, \theta_i(t) \Big)^2 dt \qquad (3\text{-}101)$$

$$= \sum_{i=1}^{N} \sum_{i=1}^{N} a_i\, a_j (\theta_1, \theta_2)_2$$

subject to the constraint that

$$\int_0^{T} f^2(t)\,dt = \sum_{i=1}^{N} a_i^{\,2} = 1 \qquad (3\text{-}102)$$

or in matrix notation, we seek

$$\eta = \max A^T B A \quad \text{with } AA^T = 1 \qquad (3\text{-}103)$$

which is given by the largest eigenvalue of B, and A is the eigenvector corresponding to the largest eigenvalue.

For comparison, we take the φ's and the channel in example 3-3, and compare the η in eq. (3-103) with the best possible ratio found in Chapter III.

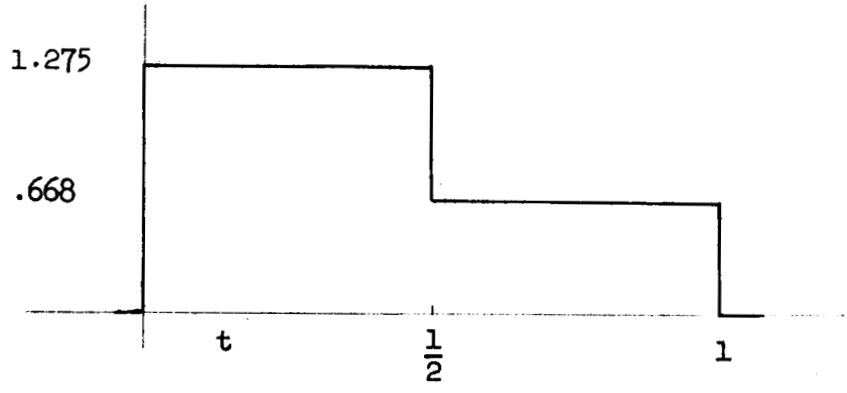
Example 3-4

Here we take $(x,y)_1 = (x,y)_2 = \int_0^1 x(t)\,u(t)\,dt$. We find

$$B = [(\theta_i, \theta_j)] = \begin{bmatrix} .168 & .048 \\ .048 & .046 \end{bmatrix}$$

and its largest eigenvalue, $\lambda_1 = 0.185$ with corresponding eigenvector, $A = [.943, .332]$. In Chapter III we found for this case that the maximum

ratio of output energy to input energy was about 0.195, so that the input signal shown in Fig. 3-6 is about 95% as good as the optimum signal.



Input Signal

Fig. 3- 6

Summary:

In this chapter, methods were developed for constructing input signals from a set of arbitrary waveforms so that after passing through an arbitrary linear channel, the resulting output signals have a prescribed inner product matrix. The innner products used for the input and output signals may be different.

It should be noted that for the computation of $(\theta_i, \theta_j) = \int_T \theta_i(t)\theta_j(t)dt$ it is not necessary to literally implement the integral of the product of θ_i and θ_j. Time sampling of the output signals is the more common procedure and is used to facilitate use in systems employing digital computers. This sampling technique is, of course, an approximation. Some approximation problems are considered in the next chapter.

Chapter 4

WAVEFORM APPROXIMATION

4.1 INTRODUCTION

In the previous chapter we dealt with waveforms which were constructed by forming a linear combination of other waveforms. There was no approximation involved in the methods developed there except the inherent approximations of potentiometer settings, amplifier gains, etc., which are inherent in any physical implementation. In this chapter we study some aspects of the problem of approximating an unknown signal by a finite sum of known signals. In several adaptive communication systems, (e.g. Glazer [15], Janowitz and White [16]), the form of the received signal is estimated by finding its projection onto a known set of signals. The measurement of these coordinates (projections) is invariably corrupted by noise, but by making a sufficient number of observations, the effect of the noise may be made arbitrarily small. A factor to be considered in such a scheme is the number of coordinates necessary to obtain a sufficiently good estimate of the received signal, as in general the smaller the number of required coordinates, the easier it is to obtain good estimates. It is of interest then to study the problem of efficient signal representation (approximation) and its implications in overall communications system design. Even when the received waveforms are known, it is usually more convenient to operate on some ordered N-tuple representation of the signals. That is, the receiver performs discrete operations on a set of numbers representing the analog signal. For example, the numbers may be the time samples of the received waveforms. If the received signal is x, and its N-tuple representation is $\hat{x} = [x_1, x_2, \ldots, x_N]$, two receivers might be built;

one operating on the analog signal x, and the other a discrete receiver operating on $\hat{x}$. For the two receivers to be equivalent, the decisions announced by each receiver must be the same. If by knowing $\hat{x}$, x (the analog signal) may be reconstructed, the discrete receiver (operating on $\hat{x}$) will necessarily make the decision as the analog receiver (operating on x). This however is not usually the case, and the discrete receiver is an approximation to the analog receiver. Although the discrete receiver may be optimum (say in the Bayes sense) for $\hat{x}$, it is not, in general, optimum for the analog signal x unless x is completely characterized by $\hat{x}$. For a given N, the choice of coordinates $x_1, \ldots, x_N$, then, determines the analog waveforms for which the discrete receiver is optimum. If the received signal is not completely characterized* by $\hat{x}$, one must assume that the representation is "close enough" so that the decisions made by the discrete receiver agree well with the decisions which would have been made by the optimum analog receiver. For a given N, and a given set of transmitted signals, performance of the discrete receiver may be enhanced by proper selection of the coordinates so that the representation is as close as possible.

In this chapter we attempt to make clear the implications and difficulties associated with the problem of finding "optimum" finite dimensional representations for given classes of signals. Although the intuitive idea of approximating "sufficiently well" any member of a given class of functions by a finite linear combination of other functions seems reasonable enough, actually

---

* By "completely characterized" is meant that for two analog signals x(t), y(t) and their discrete representations $\hat{x} = [x_1, \ldots, x_N]$, $\hat{y} = [y_1, \ldots, y_N]$, the quantities $\int_T x^2(t)dt$, $\int_T y^2(t)dt$, $\int_T x(t)y(t)dt$, must be equal to their discrete counterparts $\sum_{i=1}^{N} x_i^2$, $\sum_{i=1}^{N} y_i^2$, $\sum_{I+1}^{N} x_i y_i$.

finding these "other functions" or even the largest error incurred by the approximation where the problem is precisely stated seems to be beyond our reach. In this chapter, we develop bounds on the largest error incurred by approximating a finite set of signals by any lower dimensional subspace, thus allowing a quantitative comparison of the efficiency with which different sets of signals (chosen e.g. for their simplicity) approximates the original finite set of signals. Of particular importance here is the development of error expressions. Due to the difficulty of actually finding best finite sets of approximating functions, of particular importance in this chapter is the development of remarkably simple error expressions for exponential function approximation. Here exact expressions are developed for the smallest error incurred by approximating a function composed of a finite (or infinite) number of exponential functions by an element of the subspace spanned by a finite number of other exponentials. The form of these expressions is such that rapid trial and error calculations may be made to obtain approximate values of the best exponents of the set of exponential approximants. Finally, we examine the significance of the number of signal coordinates used in conditional maximum likelihood estimation of signal waveforms.

4.2 Models and Approximation

In engineering problems we almost invariably work with mathematical models of physical systems rather than the actual systems. Hence we deal with an approximation problem. We would like our model to approximate the actual system in the sense that the predicted performance (based on analysis of the model) agrees well, in some engineering sense, with the actual performance. It is a rare case when one can give a quantitative measure of the goodness of the

approximation, since on one hand we have a mathematical model, and on the other a physical system. We do not usually know all the parameters of the physical system, much less the exact effect of changes in these parameters on the system performance, so that a precise mathematical measure of the goodness of the approximation is impossible. If our assumptions of gaussian statistics, linearity, band-limited signals, etc., allow us to make "reasonably accurate" predictions of the performance of the actual system, then they are justified; not because the noise does have gaussian statistics, the channel is linear, the signals are band-limited, etc., but rather that the end result is satisfactory.

4.3 Mathematical Approximation Theory

The area of mathematics known as Approximation Theory as presented for example in the books by Achieser [8], Jackson [17], Korovkin [18] and Golomb [19] is a relatively new area; most of the work having been done since the turn of the century. It is currently a very active (and difficult) area of mathematical research. Only a very small portion of the problems and results from this area that are pertinent to the signal approximation problem will be discussed here. To avoid cluttering up the discussion, some of the definitions and theorems used here are collected in the appendix and are referred to by [AK] meaning the k^{th} section of the appendix.

The main problem in the theory of approximation according to Achieser can be stated as follows: "Let us suppose that two functions $f(P)$ and $F(P; A_1 \ldots A_n)$ of the point $P \in \beta$ are defined within a point set β in a space of any number of dimensions. Here $F(P; A_1 \ldots A_n)$ depends on a certain number of the parameters A_1. It is required to so determine the parameters that the deviation (or distance) of the function $F(P; A_1 \ldots A_n)$ from the function $f(P)$ for all P in β shall be a minimum." This problem is quite broad and includes

those problems for which F is a nonlinear function of the parameters A_i, and any metric [A1] may be used as the measurement of the distance D[f,F] between f and F. Among the metric spaces, the so-called normed linear spaces [A2] are important in approximation theory. If x is an element of a normed linear space, its distance from the origin is called the norm of x and is denoted by $||x||$. A normed linear space is a metric space if we put $D[x,y] = ||x-y||$. For example, the collection of all continuous functions x on $0 \leq t \leq T$ is a normed linear space if we take $||x||_c = \max |x(t)|$. Also any L^p space [A3] ($p \geq 1$) is a normed linear space in which the elements are functions x(t) on $a \leq t \leq b$ with the norm defined by

$$||x|| = \left[\int_a^b |x(t)|^p \, dt\right]^{\frac{1}{p}}.$$

The Fundamental Theorem of Approximation Theory in Normed Linear Spaces can be stated as follows (Achieser [8]): Let E be a normed linear space, and let $g_1, \ldots, g_n$ be n linearly independent elements of E. Then given $x \in E$ there exists numbers $\lambda_1, \ldots, \lambda_n$ for which the quantity

$$||x - \lambda_1, g_1 - \lambda_2, g_2 - \cdots - \lambda_n g_n||$$

attains its smallest value.

Note that the norm is not "tied down" to any particular distance measuring function. If the norm is $||x||_c$, this is the so-called Tschebycheff problem. In any approximation problem there arises two questions. First, does there exist a best approximant?" Second, "is it unique?" For the Tschebycheff problem, the above theorem says that a best approximant exists. However, the best approximant is in general not unique (see [A4]). This fact, coupled with the computational difficulty of actually finding the λ_i, makes $||\ ||_c$ an unattractive distance measuring function for the Signal Approximation Problem.

If in a linear space we can assign a number (x,y) for every pair of elements x,y called the inner product of x and y [A5] and if we take $||x|| = (x,x)^{\frac{1}{2}}$, we have what is known as a normed linear inner product space. This is what Achieser calls a Hilbert Space, although some authors reserve the name Hilbert Space for a complete [A6] normed linear inner product space.

If the approximation is in a Hilbert Space, we have the following fundamental approximation theorem [8].

Let G be a subspace of the Hilbert space H and suppose that $x \in H$ does not belong to G. If there exists in G a y whose distance from x is the shortest, then the vector x-y is orthogonal to any vector of G, i.e.

$$(x-y,\ g) = 0 \quad (g \in G)$$

By using the results of this theorem, the function

$$y = \lambda_1 g_1 + \cdots \lambda_n g_n$$

which deviates least from a given x can be presented explicitly for the case where G is generated by the linearly independent functions $g_1, \cdots g_n$. In this case [8]

$$\min_{\lambda_i} ||x-\lambda_1 g_1 - \cdots - \lambda_n g_n||^2 = \frac{G(x,g_1, \cdots, g_n)}{G(g_1, \cdots, g_n)}$$

where

$$G(g_1, \cdots g_n) = \begin{vmatrix} (g_1,g_1) & \cdots & (g_n,g_1) \\ \vdots & & \\ (g_1,g_n) & \cdots & (g_n,g_n) \end{vmatrix}$$

and $G(g_1, \cdots g_n)$ is called Gram's determinant of the functions $g_1, \cdots g_n$. The norm is "tied down" only by the fact that it is derived from an inner product, and that

$$||x+y||^2 + ||x-y||^2 = 2||x||^2 + 2||y||^2.$$

(This identity actually characterizes the inner product spaces among the normed linear spaces). In signal approximation, the approximation is almost invariably in this sense. Moreover the "distance" is usually measured in the sense of the L^2 norm, $||x||^2 = \int_T x^2(t)\,dt$. This is partly because the L^2 norm has the physically meaningful interpretation of energy, and partly due to the "goodness" of this error criterion. In 196, E.A. Guillemin's correspondence note [20] on "what is nature's error criterion" provoked some heated correspondence. Making small the integral of the square of the difference between two signals is not necessarily the same as saying that the two signals "look alike" or "sound alike", although the converse comes closer to being true. For detection problems, the "energy" of the received waveform is the important quantity, and hence the integral-squared (or L^2 norm) error is a reasonable criterion. In other applications, such as visual recognition, a more meaningful criterion is the minimization of the maximum difference between the two signals, i.e. attempt to make the signals look alike. If the signals are sufficiently smooth, an error criterion taking the smoothness into account may be used by taking the inner product to be $(x,y)_T = (x,y) + (\dot{x},\dot{y}) + \ldots + (x^{(r)}, y^{(r)})$ if the signals have r continuous derivatives. This type of approximation may be readily handled using the Fundamental Approximation Theorem in Hilbert Space, although the physical interpretation of the error is not as clear.

The usual approximation problem, as outlined above, is to choose a linearly independent set of functions $g_1, \ldots g_N$ and find the coefficients $\lambda_1, \ldots \lambda_n$ so that for a given signal x, the error $||x-\lambda_1, g_1 - \ldots - \lambda_n g_n||$ is minimized. If the signal to be approximated comes from a known class of signals x, the "goodness" of the set of approximants $g_1, \ldots g_n$, might be measured by

$$\eta_N = \max_{x \epsilon x} \min_{\lambda} ||x-\lambda_1, g_1 - \ldots - \lambda_n g_n||.$$

Different sets of N approximating functions may be compared by the value of η_N achieved. To carry the problem further, the smallest possible η_N may be sought over all N-dimensional subspaces. This last statement does not really make sense unless one specifies the class of functions from which the N approximating functions may be taken. As reasonable as it may seem , the best (in the above sense) approximating functions $g_1, \ldots, g_N$, are not necessarily elements of the set of functions to be approximated.

In contrast to the abundance of results on approximation with a given set of approximation functions, the problem of finding the best set of approximating functions has hardly been touched. The existence and uniqueness of a best set of functions $g_1 \cdots g_N$ for approximating functions of a given class has not yet been studied [19]. However in special cases, the smallest value of η_N in Eq. (A) has been calculated (see [19] p. 262).

We present this discussion so that the engineering Signal Approximation Problem may be placed in proper perspective. Almost invariably, engineering approximation is in the sense of linear approximation in Hilbert space as outlined above. Moreover, the norm used is that of L^2. In this sense, we could, in most cases, discard the words and symbols of function spaces and return to the less sophisticated sounding "integral-squared error" criterion without losing a thing. However, the notations of $||x||$, (x,y), etc., are attractive if for nothing more than ease of writing. Actually, results obtained using $\int_T [x(t) - y(t)]^2 dt$ as the criterion of error disquised as $||x-y||^2$, may hold for any inner product space, and the more general formulation may be justified.

4.4 Finite Dimensional Signal Representation

The basic idea of finite dimensional signal representation is to attempt

to characterize a signal x by an ordered N-tuple $[a_1, a_2, \cdots a_N]$. That is, knowing the numbers a_i is equivalent to knowing x, and vice versa. Unless the space X from which x is taken is N dimensional, this one-to-one correspondance cannot be obtained. A space of functions, X, is said to be N-dimensional if every $x \in X$ may be written as

$$x(t) = \sum_{i=1}^{N} C_i \, f_i(t) \qquad (4\text{-}1)$$

and N is the smallest integer for which this is true. The space X is said to be generated by the N linearly independent functions $f_1, f_2, \cdots f_N$. In general, the set of signals we wish to characterize in a particular communication system is not finite dimensional (or it may be finite dimensional, but we do not know what the f_i's are in Eq. (4-1). For a given linearly independent set of function $g_1, \cdots g_N$, we may approximate a signal x by

$$x_N(t) \approx \sum_{i=1}^{N} d_i \, g_i(t) \; . \qquad (4\text{-}2)$$

The approximation error (or the distance between x and x_N) may be measured by any metric, but since we know that at the least, the signals have finite energy, we usually measure the approximation error in the sense of the L^2 norm, $||x||^2 = \int_T x^2(t)\, dt$, and assign an inner product $(x,y) : \int_T x(t)\, y(t)\, dt$ to any two functions x and y. This leads to probably the simplest formulation of the approximation problem. If in (4-2) the g_i's are orthonormalized yielding $\varphi_1, \varphi_2 \cdots \varphi_N$, the minimization of

$$\epsilon^2 = ||x - x_N||^2 \qquad (4\text{-}3)$$

is achieved by taking $a_i = (x,\varphi_i)$, where

$$x_N(t) = \sum^{N} a_i \varphi_i(t) \ . \qquad (4\text{-}4)$$

This is the classic approximation problem in Hilbert Space as presented in the previous discussion. If in addition, we ask for the $\Phi^N = \{\varphi_1, \cdots \varphi_N\}$ which further minimizes Eq. (4-3), we are asking more than the mathematical approximation theory tells us. In the following, we attempt to make more precise what is meant by choosing the "best" Φ_N, or actually the best N-dimensional subspace.

4.5 Optimum Basis Functions for a Given Set of Signals

First of all, it is important to realize that dimensionality must refer to a set of functions, not of a single function. A single function forms a one point set, and is necessarily one dimensional, regardless of the way it may be decomposed. If the φ_i's in Eq. (4-4) are complete in L^2, and the x's are square integrable, then by taking N sufficiently large, the error defined by Eq. (4-3) may be made arbitrarily small. For finite N the error will not be zero. If a certain amount of error can be tolerated (e.g. if $||x-x_N||^2 \leq \epsilon_o^2$ the system can't distinguish between x and x_N) we may speak of a set of functions being "approximately N-dimensional".

Definition 1. A set of functions X is said to be N-dimensional with respect to a tolerable error ϵ_o^2 and a given set of orthonormal function $\Phi^N = \{\varphi_1, \cdots \varphi_N\}$ (we write this statement as $N[\epsilon_o, \Phi^N]$) if

$$\max_{x \in X} \frac{||x-x_N||^2}{||x||^2} \leq \epsilon_o^2, \qquad (4\text{-}5)$$

and if N is replaced by N-1, there exists $x_o \epsilon X$ such that

$$\frac{||X_o - X_{N-1}||^2}{||X_o||^2} > \epsilon_o^2 \qquad (4\text{-}6)$$

While this definition is (perhaps) mathematically satisfying, we would find it difficult to apply in practice since we can usually observe only a finite number of the elements of X (the set of possible signals). The complexity of the problems leaps several orders of magnitude when we attempt to find a best Φ, and the dimensionality of a set of functions with respect to only the tolerable error. Note that we are not concerned with an "average" error. This can be done, of course, by assigning some probability distribution to the set X, and seeking the Φ^N which minimizes the expected value of the integral squared error. In this case the φ_i are the eigenfunctions of the integral equation

$$\lambda\varphi(t) = \int_T R(t,\tau)\ \varphi(\tau)\ d\tau.$$

This is a corollary to the Karhunen-Loeve expansion that is usually attributed to Brown [21] in 1960, but in fact was apparently derived first by Koschmann [22] in 1954. A difficulty with such a criterion is that the representation depends upon the probability law. Also this criterion may permit large errors while minimizing the average error. Another detraction from such a random criterion is that all sets (or ensembles) having the same $R(t,\tau)$ yield the same representation function without regard to the actual time waveforms. For example, the random telegraph signal (a random square wave assumingvales of -1 or 1 with equal probability, and the probability that K amplitude changes occur in a time interval of length T is given by the Poisson distribution), and the output of a law pass RC circuit due to white noise, both would have

the same representation functions even though their time waveforms are are completely different. As is often the case in communication theory, the assignment of a probability law allows us to solve a problem, but is not necessarily the problem we started out with.

Definition 2. A set of functions X is said to be N-dimensional with respect to a tolerable error ϵ_o^2 (written as $N[\epsilon_o]$) if

$$N[\epsilon_o] = \min_{\Phi^N} N\left[\epsilon_o, \Phi^N\right] \tag{4-7}$$

That is, we seek the smallest number N, and the associated Φ^N (or $\Phi^{N's}$) for which equation (4-5) and (4-6) are satisfied. The functions φ_1 (since they are orthonormal) are constrained only to lie in L^2. The complexity of this problem should be apparent. The search for the best Φ^N is equivalent to fixing N and seeking those functions φ_i which achieve

$$\min_{\Phi^N} \max_{x\epsilon X} \left|\left| x - \sum_{i=1}^{N} (x_1\ \varphi_i)\ \varphi_i \right|\right|^2. \tag{4-8}$$

There are no theorems in approximation theory on which we can draw in order to aid in the solution of this problem, and yet it is just this kind of problem that is implied when one intuitively speaks of a set of functions being "approximately N-dimensional".

It is the more remarkable then, that Slepian, Landau, and Pollack ([23]) [24] have succeeded in obtaining results on the "best" Φ^N for a particular class of functions, and lend preciseness to the intuitive notion that a signal may be characterized by approximately 2WT numbers if the bandwidth of the signal is "about" W, and its time duration is "about" T. This notion is based

on the well known "sampling theorem" which states that a strictly bandlimited signal f. of bandwidth W may be characterized by its time samples spaced 1/2W seconds apart, and the representation takes the form

$$f = \frac{1}{2W} \sum_{-\infty}^{\infty} f\left(\frac{n}{2W}\right) \operatorname{sinc} \frac{2\left[t - \frac{n}{2W}\right]}{2\left[t - \frac{n}{2W}\right]} . \qquad (4\text{-}9)$$

This has led a number of misuses of this theorem. One such misuse is the notion "That if f(t) is small outside $\frac{-T}{2} < t < \frac{T}{2}$, then of course one only need take 2WT samples". Another notion is that "if f is bandlimited to W, and time limited to T, the f may be characterized (uniquely in fact) by 2WT samples". Actually this latter notion is true (vacuously) since the class of functions which are strictly bandlimited, and strictly time-limited is an empty set.

The class of functions considered by Pollack and Tondau [24], is the class of strictly bandlimited functions of bandwidth W, whose energy outside $-T/2 \leq t \leq T/2$ is <u>equal</u> to ϵ_T^2 -(denoted by $E(\epsilon_T)$. They show that for this class of function, the function $\varphi_o, \varphi, - \varphi_{N-1}$ which achieve

$$\min_{\Phi^N} \max_{f \in E(\epsilon_T)} \min_{a_i} \int_{-\infty}^{\infty} \left| f(t) - \sum_{o}^{N-1} a_i \psi_i(t) \right|^2 dt \qquad (4\text{-}10)$$

are the angular prolate spheriodal function ψ_o, -- ψ_{n-1}. In regard to the difference in approximating with the sampling function rather than prolate spheriodal functions, they show that if $f \in E(\epsilon_T)$, then

$$\inf_{a_i} \int_{-\infty}^{\infty} \left| f(t) - \sum_{o}^{[2WT]+N} a_n \varphi_n \right|^2 dt < C \epsilon_T^2$$

is

(a) true for all such f with N = 0, C = 12, if the φ_n are the prolate spheriodal wave functions

(b) false for some such f for any finite constants N and C if the φ_n are sampling functions.

As important as these results certainly are, we should not misconstrue them to say that the prolate spheriodal wave functions are the best representation functions for signal approximation problems. This would be true if our signals satisfied the conditions of the above theorems. For some other class of functions the prolate spheriodal wave function may provide a very poor approximation, compared with the same number of other orthogonal signals.

The work done by W. H. Huggins, et. al. [25], [26], [27], [28], has been primarily concerned with the use of orthonormalized exponential functions. In many ways the exponential functions are as interesting a class of signals as the prolate spheriodal function or sampling function. The Gram-Shmidt orthonormalization process has a particularly simple form as shown by Katz [29]. An infinite set of exponential $\left\{\epsilon^{-a_i t}\right\}$ may be complete in L^2, (Schaz's Theorem) see [25] p _____ . No small point in their favor is that the ease of generation of these signals (see [26]) is considerably greater than that of say the prolate spheriodal wave functions. Huggins and Young demonstrate in [27] that exponential function do a remarkably good job of approximating electrocardiograms. Certainly we would expect a class of electrocardiograms to have more structure to it than just that they have some "essential" time duration and "essential" bandwidth. In this connection, we note that even though exponential functions are "good" for approximating electrocardiograms the tag of "best signals" is still mathematically indefensible. A major difficulty in the search for a "best" set of basis functions for a physical ensemble of signals is in describing the ensemble. In practice one can observe

only a finite number of the members of the ensemble. The question of how well and in what sense a set of functions can be described or characterized by a finite number of its elements can possibly be answered for a well defined set of functions. For physical ensembles of signals, however, one can only assume that if enough sample signals are observed, that they contain the "essential" characteristics of the ensemble. Whether the characterization is good or bad, it is all that one can do.

4.6 Best Representation Functions for a Finite set of Signals

A set of M signals $\{f_1, f_2, \text{---} f_M\}$ is at most M dimensional. If the f_i's are linearly independent, Gram's determinant is greater than zero, hence the rank of the inner product matrix $[(f_i, f_j)]$ is M. If each f_i is generated by K linearly independent functions, the dimensionality of the set of f_i's is $\leq K < M$. It is well known [30] that the relative size of a Grams determinant is an indication of the "closeness" or "near dependance" of the f_i's. In Courant and Hilbert [30], the "measure of independence is taken to be the size of the smallest eigenvalue of the of the quadratic form

$$K(t,t) = \int_T (t_1 f_1 + \text{----} t_M f_M)^2 = \sum_i^M \sum_j^M (f_i, f_j)\, t_i\, t_k \qquad (4\text{-}11)$$

If the eigenvalues of eq. (4-11) are such that N of the M eigenvalues are "significantly larger" than the other M - N eigenvalues, it is sometimes said that the set of functions f_i, --, f_m is "essentially N dimensional". In order for this statement to have meaning, the meaning of term "essentially" must be made more precise. The "dimensionality" of a finite set of functions needs to be examined in the light of our previous definitions of $N[\epsilon_o, \Phi]$ and $N[\epsilon_o]$. Recall that for the "best" approximating functions, we seek those which provide

Rearranging the terms in (4-14), we seek

$$\max_{\Phi^N} \sum_{k=1}^{N} \int_T \int_T \sum_{i=1}^{M} \frac{f_i(t)\, f_i(\tau)}{\| f_i \|^2} \varphi_k(t)\, \varphi_k(\tau)\, dt d\tau \qquad (4\text{-}15)$$

Note that each term in the sum on K is of the forms

$$\int_T \int_T K(t,\tau)\, \varphi(t)\, \varphi(\tau)\, dt t\tau. \qquad (4\text{-}16)$$

The maximum of (4-16) is well known [3] and is given by the largest eigenvalue of the integral equation

$$\lambda\varphi(t) = \int_T K(t,\tau)\, \varphi(\tau)\, d\tau. \qquad (4\text{-}17)$$

Call this largest eigenvalue λ_1, and its corresponding eigenfunction, φ_1. The next φ, φ_2, is chosen to maximize (4-16) with the additional constraint that $(\varphi_2, \varphi_1) = 0$. This maximum is given by the next largest eigenvalue of (4-17). Continuing in this manner, we find that

$$\min_{\Phi^N} \sum_{i=1}^{M} \epsilon_i^2 = M - \sum_{k=1}^{N} \lambda_k \qquad (4\text{-}18)$$

Since $K(t,\tau) = \sum_{i=1}^{M} f_i(t)\, f_i(\tau)$, the λ_k are found as the eigenvalues of the symmetric matrix

$$G = \left[\frac{(f_i, f_j)}{\| f_i \| \cdot \| f_j \|} \right] \qquad (4\text{-}19)$$

This is the inner product matrix of the <u>normalized</u> f_i's. The eigenvectors of (4-19) are orthogonal and may be normalized.

$$\min_{\Phi^N} \max_{f_i} \min_{a_k} \frac{\left\| f_i - \sum_{k=1}^{N} a_k \varphi_k \right\|^2}{\left\| f_i \right\|^2} . \qquad (4\text{-}12)$$

$$i = 1, \text{ --- } M$$

We begin by seeking the Φ^N which provide

$$\min_{\Phi^N} \sum_{i=1}^{M} \epsilon_i^2 = \min_{\Phi^N} \sum_{i=1}^{M} \frac{\left\| f_i - \sum_{k=1}^{N} a_k \varphi_k \right\|^2}{\left\| f_i \right\|^2} \qquad (4\text{-}13)$$

It seems more reasonable to consider the normalized error in eq. (4-13), as $\left\| f - \sum_{k=1} a_k \varphi_k \right\|^2$ may be quite small while the normalized error may be considerably larger if $\left\| f \right\|^2$ is also small. The Φ^N which satisfy eq. (4-13), do not, as we will show, satisfy eq. (4-12) (in general), but the solution of this "least squares" problem does provide a figure of comparison for different Φ^N.

If the $\varphi_i = \lambda_{1_i} f_1 + \text{ -- } + \lambda_{M_i} f_M$, the problem may be considered an application of a technique in Factor Analysis called Hotelling's method, and the problem reduces to determing the coefficients λ_i -- λ_M As the problem is linear, one would expect that the best φ_i's would have this form. However, in the following this is not assumed, and the φ_i's are constrained only to be orthonormal.

Recalling that $(x,y) = \int_T x(t)\, y(t)\, dt$ and expanding (4-13), we now seek

$$\max_{\Phi^N} \sum_{i=1}^{M} \sum_{K=1}^{N} \int_T \int_T f_i(t)\, f_i(\tau)\, \varphi_k(t)\, \varphi_k(\tau)\, dt d\tau . \qquad (4\text{-}14)$$

The eigenfunctions of eq. (4-16) are given by

$$\varphi_\ell = \frac{1}{\sqrt{\lambda_\ell}} \sum_{i=1}^{M} \mu_{\ell i} f_i \tag{4-20}$$

Where the f_i's are normalized, and $[\mu_{\ell 1}, \cdots \mu_{\ell m}]$ is the ℓ^{th} eigenvector of G. Then

$$(\varphi_\ell, \varphi_k) = \frac{1}{\sqrt{\lambda_\ell \lambda_k}} \sum_{i}^{M} \sum_{j}^{M} \mu_{\ell i} \mu_{kj} (f_i, f_j) \tag{4-21}$$

$$= \frac{1}{\sqrt{\lambda_\ell \lambda_k}} \begin{bmatrix} \mu_{\ell 1} \\ \vdots \\ \mu_{\ell m} \end{bmatrix} [(f_i, f_j)] [\mu_{ki} \cdots \mu_{km}]^T \tag{4-22}$$

$$= \delta_{k\ell} \tag{4-23}$$

The sum of the squares of the distance from each of the f_i (normalized) to any N dimensional subspace is greater than or equal to the quanitity

$$M - \sum_{k=1}^{N} \lambda_k \tag{4-24}$$

Now since $\sum_{i=1}^{M} \epsilon_i^2 \geq M - \sum_{k=1}^{N} \lambda_k$ for any Φ^N,

and since the minimum of max $\{\epsilon_i^2\}$ would be attained if the ϵ_i^2 equal, if follows that for any Φ^N

$$\min_{\Phi^N} \max_{f_1 \cdots f_m} \{\epsilon_i^2\} \geq \frac{M - \sum_{k=1}^{N} \lambda_k}{M} = \Omega_N \tag{4-25}$$

The eigenfunctions given by eq. (4-20) will usually not be convenient to work with, and moreover do not usually provide the best functions required by eq. (4-22). These eigenfunctions could be considered as the "best"

representation functions in the sense that they satisfy eq. (4-23), i.e. they provide the smallest average error. Huggins and Young, [27] have made use of the eigenvectors of the unnormalized correlation matrix to "essentially" represent the original functions $f_1, \ldots f_M$. The quantity Ω_N is not claimed to be a greatest lower bound, but it does provide a figure of comparison for approximating N functions by N orthonormal functions, selected, say, for ease of construction.

The simplest example where the bound in eq. (4-25) is attained is the approximation of any two functions by a single function. The best single approximating function for a set of two functions $\{f_1, f_2\}$, with $(f_1, f_2) > 0$, is given by $\psi = \frac{f_1 + f_2}{||f_1 + f_2||}$ as may be verified directly.

Without some sort of bound such as Ω_N, one has no way of judging how well a particular set of orthonormal functions approximates a given set of signals.

Example 4-1

As an example of the determination of the "essential dimensionality" of a finite set of signals, we consider the normalized signals $\sqrt{2}\,\epsilon^{-t}$, $\sqrt{4}\,\epsilon^{-2t}$, $\sqrt{6}\,\epsilon^{-3t}$, $\sqrt{8}\,\epsilon^{-4t}$, $\sqrt{10}\,\epsilon^{-5t}$. Their innerproduct matrix (where $(x,y) = \int_0^\infty x(t)\,y(t)\,dt$) is given by

$$\begin{bmatrix} 1.0000 & 0.9428 & 0.8666 & 0.8000 & 0.7454 \\ 0.9428 & 1.0000 & 0.9798 & 0.9428 & 0.9035 \\ 0.8666 & 0.9798 & 1.0000 & 0.9897 & 0.9682 \\ 0.8000 & 0.9428 & 0.9897 & 1.0000 & 0.9938 \\ 0.7454 & 0.9035 & 0.9682 & 0.9938 & 1.0000 \end{bmatrix}$$

and its eigenvalues are given by

$$\lambda_1 = 4.6586680$$
$$\lambda_2 = 0.3239158$$
$$\lambda_3 = 0.0169404$$
$$\lambda_4 = 0.0004492$$
$$\lambda_5 = 0.0000267$$

If only the first eigenfunction is used, the maximum error is found to be 0.18987, and the set of 5 signals may be said to be one-dimensional with respect to this eigenfunction if an error of 0.18987 can be tolerated. If N = 2 and the first two eigenfunctions are used as the representative function, the maximum errror is found to be 0.005599, and for N = 3 the maximum error is 0.00016556. In any case the dimension ascribed to this set of five functions depends upon the tolerable error. If it is required, for example, that the maximum representation errror be less than 10^{-6}, the dimension of this set would be 5, since the bound

$$\Omega_4 = \frac{5 - \sum_{i=1}^{4} \lambda_i}{5} = 5.3 \times 10^{-6}$$

In the next example, we examine the "goodness" of the approximation of three signals by two functions which are the best of their class, and compare the maximum error with the bound Ω_2, and the errors obtained by using the first two eigenfunctions of eq. (4-17).

Example 4-2

Here we take M = 3 and the three normalized signals to be $\sqrt{2a_1}\,\epsilon^{-\alpha_1 t}$, $\sqrt{8}\,\epsilon^{-4t}$, $\sqrt{2a_2}\,\epsilon^{-\alpha_2 t}$ where $\alpha_1 = (25 - \sqrt{36a})4$ and $\alpha_2 = (25 + \sqrt{36a})4$. The inner product matrix is

$$\begin{bmatrix} 1.0000000 & 0.8834522 & 0.6400000 \\ 0.8834522 & 1.0000000 & 0.8834522 \\ 0.6400000 & 0.8834522 & 1.0000000 \end{bmatrix}$$

with eigenvalues

$$\lambda_1 = 2.609719$$
$$\lambda_2 = 0.360000$$
$$\lambda_3 = 0.0302808.$$

For $N = 2$, $\Omega_2 = (3 - \sum_{i=1}^{2} \lambda_i)/3 = 0.0101$. That is, for any two-dimensional representation, the maximum error must be ≥ 0.0101. If the first two eigenfunctions of eq. (4-17) are used as representation functions, we find the errors to be

$$\epsilon_1^2 = 0.005692$$
$$\epsilon_2^2 = 0.018897$$
$$\epsilon_3^2 = 0.005692.$$

We now compare these results with those obtained by finding the representation error using the two-dimensional subspace spanned by the functions $\epsilon^{-\alpha t}$, $\epsilon^{-\beta t}$. We will show that $\alpha = 2$, $\beta = 8$ provide the best two-exponential subspace. If we denote by $S[\epsilon^{-\alpha t}, \epsilon^{-\beta t}]$ the subspace spanned by $\epsilon^{-\alpha t}$ and $\epsilon^{-\beta t}$, and by $||f - S[\epsilon^{-\alpha t}, \epsilon^{-\beta t}]||$ the distance from f to S, it turns out that for $\alpha = 2$, $\beta = 8$,

$$\max_{f_i} \frac{||f_i - S[\epsilon^{-2t}, \epsilon^{-8t}]||^2}{||f_i||^2} = \frac{1}{81} = 0.0123. \qquad (4\text{-}26)$$

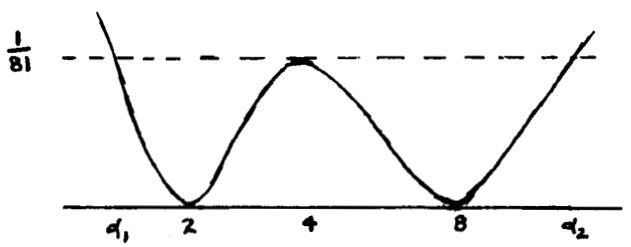
This is to be compared with the maximum error of 0.0189 obtained using the eigenfunctions of eq. (4-17) as representation functions. The maximum error

obtained using the subspace spanned by ϵ^{-2t} and ϵ^{-8t} is also quite close to the lower bound $\Omega_2 = 0.0101$ so that in a very real sense these two simple exponentials provide a good approximation. The remainder of this exponential approximations.

We first make use of a result derived later in this chapter that the representation error obtained by approximating the normalized function $\sqrt{2a}\ \epsilon^{-\alpha t}$ by $S\ [\epsilon^{-\alpha t}, \epsilon^{-\beta t}]$ is given by

$$\epsilon^2 = \left(\frac{a-\alpha}{a+\alpha}\right)^2 \left(\frac{a-\beta}{a+\beta}\right)^2 \qquad (4\text{-}27)$$

For $\alpha = 2$, $\beta = 8$, the errors obtained by setting $a = (25-\sqrt{369})/4$, 4, $(25+\sqrt{369})/4$ are all equal and have the value $\frac{1}{\delta_1}$. This equal error property does not by itself necessarily imply that $\alpha = 2$, $\beta = 8$ are the best exponents. However, over the interval $(25-\sqrt{369})/4 \leq a \leq (25+\sqrt{369})/4$, the equation $\frac{\partial \epsilon^2}{\partial a} = 0$ is satisfied for $a = 2, 4, 8$. $a = 4$ provides the only maximum. This is indicated by the sketch in Fig. (4-1). It is seen that any perturbations about $\alpha = 2$, $\beta = 8$ must result in a maximum error exceeding $\frac{1}{\delta_1}$.



ϵ^2 as a function of the parameter a

Fig. 4-1

Even though the functions ϵ^{-2t}, ϵ^{-8t} provide remarkably good approximants to the three original signals, we have not shown that they are <u>the</u> best functions to use. We do know how well they perform compared to the bound Ω_2

(which may be too small). In practice, this is all one can do; select a set of N representation function, and by actual trial, determine if the representation is satisfactory. We would also like to have an idea as to how much to expect from this N-dimensional representation, and the bound in eq. (4-25) provides a quantitative measure of the goodness of the approximation.

Considerable work has been done using sets of expoential functions as approximating functions. This has been done primarily and almost exclusively at Johns Hopkins University. Of particular relevance here is the work by McDonough [28] who attempts to find the best exponents α_i for the functions $\epsilon^{-\alpha_1 t}, \epsilon^{-\alpha_N t}$ to approximate a given signal. This has also been done by A.A. Wolfe (unpublished) in connection with an adaptive communication system. An approach to their problem is to take a sufficient number of the functions $\epsilon^{-\alpha_i t}$ so that the representation error is considerably smaller than the allowable error, and then use some perturbation scheme to reduce the number of representation function required to achieve the required error. The problem is mathematically straightforward (for approximating a single function) but a practical difficulty arises in that the approximation error has been noted to be relatively insensitive to perturbations of the α_i, and thus finding the best α_i is computationally very difficult.

In the following, we develop some rather remarkably simple and computationally useful error expressions for approximation using exponential functions.

4.7 Error Expressions for Exponencial Approximation

In order to appreciate the simplicity and usefulness of the results to follow, we examine the standard procedures for finding the distance of a function x from the linear subspace spanned by N linearly independent functions $g_1, \cdots g_N$. A fundamental result in approximation theory (ACHIESER [8] p. 15) shows that

$$\epsilon^2 = \min_{a_i} || x - a_1 g_1 - a_2 g_2 \cdots - a_N g_N ||^2 = \frac{G(x, g_1, \cdots g_N)}{G(g_1, \cdots g_N)} \qquad (4\text{-}28)$$

where

$$G(g_1, \cdots g_N) = \begin{bmatrix} (g_1, g_1) & (g_2, g_1) \cdots (g_N, g_1) \\ (g_1, g_2) & (g_2, g_2) \cdots (g_N, g_2) \\ \hline (g_1, g_N) & (g_2, g_N) \cdots (g_N, g_N) \end{bmatrix} \qquad (4\text{-}29)$$

Of course the set of functions $g_1, \cdots g_N$ may be orthonormalized yielding $\varphi_1, \varphi_2 \cdots \varphi_N$, in which case the squared error is given by

$$\epsilon^2 = || x ||^2 - \sum_{i=1}^{N} (x, \varphi_i)^2. \qquad (4\text{-}30)$$

The simplicity of this expression is only apparent, since the expression for the φ's in terms of the g's is (see previous chapter) just as involved as (4-28) and in fact the same amount of labor is involved whether one uses Eq. (4-28) or Eq. (4-30). Also it, of course, makes no difference how the φ's are obtained from the g's. The space spanned by the φ's is the same as that spanned by the g's. Equation (4-30) has the conceptual advantage of exhibiting the decreasing in error if the number of φ's is increased. In any case, equations (4-29) and (4-30) are in their simplest form. That is, they cannot, in general, be arranged to exhibit the part of the error due to use of a particular g. In particular, if the g's have the form $g(t:\theta_i)$, the effect on the approximation error of perturbations of the parameters θ_i can usually be obtained only by carrying out the operation indicated in Eq. (4-29) for different values of the parameters and observing the result. In general, changing the value of a single parameter would require a complete recomputation of the error.

4.8 Single Exponential Approximated by a Set of Exponentials

We derive here an expression for the squared error

$$\epsilon^2 = \min_{a_i} \frac{|| \epsilon^{-at} - \alpha_1 \epsilon^{-a_2 t} \ldots \alpha_N \epsilon^{-a_N t} ||^2}{|| \epsilon^{-at} ||^2} . \qquad (4\text{-}31)$$

that is a single exponential, ϵ^{-at}, is to be approximated by a linear combination of the functions $\epsilon^{-a_1 t}, \ldots, \epsilon^{-a_N t}$. First we compute the error of approximating a single normalized exponential, $f = \sqrt{2a}\epsilon^{-at}$, by another single normalized expnential functions, $g_i = \sqrt{2a_i}\, \epsilon^{-a_i t}$.

By direct computation, the error is given by

$$\epsilon_i^2 || f ||^2 - (f, g_i)^2$$

$$\epsilon_i^2 = 1 - \frac{4aa_i}{(a+a_i)^2} \qquad (4\text{-}32)$$

$$\epsilon_i^2 = \left(\frac{a-a_i}{a+a_i} \right)^2$$

We now show that Eq. (4-31) reduces to

$$\epsilon^2 = \min_{a_i} \frac{|| \epsilon^{-at} - \alpha_1 \epsilon^{-a_1 t} \ldots \alpha_N \epsilon^{-a_N t} ||^2}{|| \epsilon^{-at} ||^2} = \epsilon_1^2 \cdot \epsilon_2^2 \cdot \ldots \epsilon_N^2 \qquad (4\text{-}33)$$

The result is remarkably simple. The effect of adding another expoential $\epsilon^{-a_{N+1}}$ to the set of approximants $\epsilon^{-a_1 t} \ldots \epsilon^{-a_N t}$, is to <u>multiply</u> ϵ_{N+1}^2 by the error already obtained. Also, the effect of varying a particular a_i is clearly exhibited, and one only has to recompute the single term c_i^2. Note that ϵ_i^2 is the normalized error of approximating ϵ^{-at} by $\epsilon^{-a_i t}$, and the total error is the product of these individual errors. This is certainly not true for arbitrary functions.

The proof of Eq. (4-33) is readily obtained using a result by Achieser [8] on approximating the function t^q by a linear combination of the functions $t^{p_1}, t^{p_2} \ldots t^{-p_N}$. A simpler proof can be obtained following the outlines of the proof we use later on the approximation of a sum of exponentials. Acheiser's result is that

$$\epsilon^2 = \frac{G(t^q, t^{p_1}, \ldots t^{p_N})}{G(t^{p_1}, \ldots t^{p_N})} = \frac{1}{2q+1} \prod_{i=1}^{N} \left| \frac{q-p_i}{q+p_i+1} \right|^2 \tag{4-34}$$

where $(t^\alpha, t^\beta) = \int_0^1 t^\alpha t^\beta \, dt$. By making the change of variable $t = \epsilon^{-x}$, it is easy to obtain Eq. (4-33) from Eq. (4-34).

4.9 A Sum of Exponentials Approximated by a Set of Exponentials

While the above result is both interesting and useful, a much more practically significant result would be a simple, computationally useful expression for the error incurred by approximating a sum of exponentials,

$$f = \sum_{i}^{M} \beta_i \epsilon^{-\alpha_i t} \tag{4-35}$$

by a linear combination of others exponentials $\epsilon^{-a_1 t}, \ldots \epsilon^{-a_N t}$. As was remarked earlier, we may use Eq. (4-37) as an approximation to a given function or set of functions, and by taking M sufficiently large, the approximation error may be made as small as desired. We would then try to chose the N a_i's (N < M) in order to achieve a prescribed allowable error.*

A simple error expression for the approximation of a sum of exponentials can in fact be obtained, but is apparently somewhat harder to show.

---

* This procedure is justified since if f is an approximant to g and $\hat{f}$ is an approximant to f, than by the triangle inequality,

$$||g-\hat{f}||^2 = ||g-f+f-\hat{f}||^2 \leq [\,||g-f|| + ||f-\hat{f}||\,]^2.$$

Roughly speaking, this says that if f is close to g and $\hat{f}$ is close to f, then $\hat{f}$ is close to g.

First, we suppose that the set of approximating functions $\epsilon^{-a_1 t}$, $\epsilon^{-a_2 t}$ ... $\epsilon^{-a_N t}$ are orthonormalized in any manner yielding φ_1, ... φ_N. We wish to approximate a function

$$f = \sum_{i}^{M} \beta_i \epsilon^{-\alpha_i t} = \sum_{i}^{M} \beta_i f_i(t) \tag{4-36}$$

(We may consider that the subspace spanned by $\epsilon^{-a_1 t}$, ... $\epsilon^{-a_M t}$ is arbitrarily close to a set of functions we are attempting to characterize).

The approximation error is given by Eq. (4-30)

$$\epsilon^2 = || f ||^2 - \sum_{k=1}^{N} (f, \varphi_k)^2 \tag{4-37}$$

$$= || \sum_{i=1}^{M} \beta_i f_i ||^2 - \sum_{k=1}^{N} \left[\sum_{i=1}^{M} \beta_i (f_i \varphi_k) \right]^2 \tag{4-38}$$

$$= \sum_{i=1}^{M} \sum_{j=1}^{M} \beta_i B_j (f_i f_j) - \sum_{k=1}^{N} \sum_{i=1}^{M} \sum_{j=1}^{M} \beta_i \beta_j (f_i \varphi_k)(f_j \varphi_k) \tag{4-39}$$

By rearranging the terms of this expression, we may write

$$\epsilon^2 = \left\{ \beta_1^2 || f_1 ||^2 - \sum_{k=1}^{N} (f_1 \varphi_k)^2 \right\} + \beta_2^2 \left\{ || f_2 ||^2 - \sum_{k=1}^{N} (f_2 \varphi_k)^2 \right\} + \cdots$$

$$+ \cdots + \beta_M^2 \left\{ || f_M ||^2 - \sum_{k=1}^{N} (f_M \varphi_k)^2 \right\}$$

$$+ 2\beta_1 \beta_2 \left\{ (f_1 f_2) - \sum_{k=1}^{N} (f_1 \varphi_k)(f_2 \varphi_k) \right\} + 2\beta_1 \beta_3 \left\{ (f_1 f_3) - \sum_{k=1}^{N} (f_1 \varphi_k)(f_3 \varphi_k) \right\} + \cdots$$

$$+ \cdots + 2\beta_{M-1} \beta_M \left\{ (f_{M-1} f_M) - \sum_{k=1}^{N} (f_{M-1} \varphi_k)(f_M \varphi_k) \right\} . \tag{4-40}$$

So far, we have not made use of the exponential nature of the functions, i.e., Eq. (4-40) holds for the approximation of the sum of arbitrary functions f_i by arbitrary orthonormal functions φ_k. We note that Eq. (4-40) involves only the coefficients in Eq. (4-36), the ϵ_i^2, and terms of the form

$$\left[(f_a f_b) - \sum_{k=1}^{N} (f_a \varphi_k)(f_b \varphi_k)\right] \tag{4-41}$$

which may be written as

$$(f_a - \sum^{N} (f_a \varphi_k)\, \varphi_k,\ f_b - \sum^{N} (f_b \varphi_k)\, \varphi_k) \tag{4-42}$$

then using the Schwartz inequality

$$\left| (f_a - \sum^{N} (f_a \varphi_k)\, \varphi_k,\ f_b - \sum^{N} (f_a \varphi_k)\, \varphi_k) \right| \leq \sqrt{\epsilon_a^2 \epsilon_b^2}\ . \tag{4-43}$$

We have not yet made use of the exponential nature of the functions. We show if the functions are exponentials, that

$$(f_a f_b) - \sum_{k=1}^{N} (f_b \varphi_k)(f_a, \varphi_k) = (f_a, f_b)\, \epsilon_a \epsilon_b \tag{4-44}$$

where $\epsilon_a = \prod_{i=1}^{N} \left(\frac{a - a_i}{a + a_i}\right)$, $\epsilon_b = \prod_{i=1}^{N} \left(\frac{b - a_i}{b + a_i}\right)$, and

hence eq. (4-40) may be written as

$$\epsilon^2 = \sum_{i}^{M} \sum_{j}^{M} \beta_i \beta_j\, \epsilon_i \epsilon_j (f_i f_j), \tag{4-45}$$

a qradratic form well suited for hand or machine computation. Also the effect of varying the a_i's is clearly exhibited.

The remainder of this section will be devoted to proving: If $f_a = \epsilon^{-at}$ $f_b = \epsilon^{-bt}$, and the φ_k are obtained by orthonormalizing the set of functions $\epsilon^{-a_1 t}, \ldots, \epsilon^{-a_N t}$,

$$\text{then} \qquad \left(f_a, f_b\right) - \sum^{N} (f_a \varphi_k) = (f_a, f_b)\, \epsilon_a \epsilon_b \tag{4-46}$$

$$\text{where} \qquad \epsilon_a = \prod_{i=1}^{N} \left(\frac{a - a_i}{a + a_i} \right) \text{ and } \epsilon_b = \prod_{i=1}^{N} \left(\frac{b - a_i}{b + a_i} \right), \tag{4-47}$$

$$\text{and } (x,y) = \int_0^{\infty} x(t)\, y(t)\, dt.$$

The means by which the φ_k are obtained from the linearly independent set of functions $\epsilon^{-a_1 t}, \ldots, \epsilon^{-a_N t}$ is of course immaterial. Whatever orthonormalization technique is used, the φ_k will have the general form

$$\varphi_k(t) = \sum_{i=1}^{N} a_{ki}\, \epsilon^{-a_i t}. \tag{4-48}$$

$$\text{Then} \qquad (f_a, \varphi_k) = \sum_{i=1}^{N} a_{ki} \frac{1}{a+a_i}, \tag{4-49}$$

$$(f_b, \varphi_k) = \sum_{i=1}^{N} a_{ki} \frac{1}{b+a_i}, \tag{4-50}$$

$$\text{and} \qquad (f_a, \varphi_k)(f_b, \varphi_k) = \left[\sum_{i=1}^{N} a_{ki} \frac{1}{a+a_i} \right] \left[\sum_{i=1}^{N} a_{ki} \frac{1}{b+a_i} \right]. \tag{4-51}$$

We can write Eq. (4-51) as

$$(f_a, \varphi_k)(f_b, \varphi_k) = \frac{\rho_k(a, b; a_1, \ldots a_N)}{\prod_{i=1}^{N} (a+a_i) \prod_{i=1}^{N} (b+a_i)} \tag{4-52}$$

where ρ_k is a polynomial of degree no higher than 2N (e.e. the highest exponents of a and b are no higher than N). Carrying out the summation indicated in Eq. (4-45),

$$q \sum_{k=1}^{N} (f_a\varphi_k)(f_b\varphi_k) = \sum_{k=1}^{N} \frac{\rho_k(a, b; a_1, \cdots a_N)}{\prod_{i=1}^{N} (a+a_i) \prod_{i=1}^{N} (b+a_i)} \tag{4-53}$$

$$= \frac{P_N(a, b; a_1, \cdots a_N)}{\prod_{i=1}^{N} (a+a_i) \prod_{i=1}^{N} (b+a_i)}, \tag{4-54}$$

where $P_N = \sum_{k=1}^{N} \rho_k$ is also a polynomial of degree $\leq 2N$ since each ρ_k was of degree no higher than 2N. Then Eq. (4-45) becomes

$$\frac{1}{a+b} - \frac{P_N(a,b; a_1, \cdots a_N)}{\prod_{i=1}^{N} (a+a_i) \prod_{i=1}^{N} (b+a_i)} \tag{4-55}$$

$$= \frac{1}{a+b} \quad \frac{P'_N(a, b; a_1, \cdots a_N)}{\prod_{i=1}^{N} (a+a_i) \prod_{i=1}^{N} (b+a_i)} . \tag{4-56}$$

Now if a or b equals any of the a_i, eq. (4-55) becomes zero since the approximation error is zero (see Eq. (4-43)). This means that the polynomial P'_N is divisible by $\prod_{i=1}^{N} (a-a_i)$ and $\prod_{i=1}^{N} (b-a_i)$, or

$$P'_N = A_N \prod_{i=1}^{N} (a-a_i) \prod_{i=1}^{N} (b-a_i) \tag{4-57}$$

We have then (4-58)

$$(f_a,f_b) - \sum^{N} (f_a,\varphi_k)(f_b,\varphi_k) = (f_a,f_b)\, A_N \frac{\prod^{N} (a-a_i) \prod^{N} (b-a_i)}{\prod (a+a_i) \prod (b+b_i)} .$$

We need only to find A_N. If we let both a and b become arbitrarily large, we see from Eqs. (4-48) and (4-49) that

$$\lim_{a,b, \to \infty} (f_a\varphi_k)(f_b\varphi_k) = 0. \tag{4-59}$$

Also, we see that

$$\lim_{a,b, \to \infty} \frac{\prod^{N} (a-a_i) \prod^{N} (b-a_i)}{\prod^{N} (a+a_i) \prod^{N} (b+a_i)} \tag{4-60}$$

Thus from Eq. (4-57), $A_N = +1$ independent of N. We have shown then that

$$(f_a,f_b) - \sum_{k=1}^{N} (f_a\varphi_k)(f_b\varphi_k) = (f_af_b)\, \epsilon_a\, \epsilon_b. \tag{4-61}$$

where ϵ_a^2 and ϵ_b^2 represent the squared distance of $\sqrt{2a}\, \epsilon^{-at}$ and $\sqrt{2b}\, \epsilon^{-at}$ respectively, from the subspace spanned by the functions ϵ^{-a_1t}, ϵ^{-a_2t}, .. ϵ^{-a_Nt}. That is, ϵ_a^2 and ϵ_b^2 are the <u>normalized</u> errors.

By setting a = b, Eq. (4-33) is obtained. However, Eq. (4-61) cannot be obtained from Dq. (4-33). Note that the <u>sign</u> of ϵ_a or ϵ_b is determined by the defining Eq. (4-47) and is not $\pm\sqrt{\epsilon_a^2}$.

4.10 <u>Maximum Error of Approximating Sum of Exponentials</u>

The squared norm of a function

$$f(t) = \sum_{i=1}^{M} \beta_i\, f_i(t),\ f_i(t) = \epsilon^{-\alpha_i t} \tag{4-62}$$

is given by $||f||^2 = \sum_i^M \sum_j^M \beta_i \beta_j (f_i f_j)$. The normalized squared error of approximating a function of the form (4-42) is obtained by dividing Eq. (4-45) by $||f||^2$

$$\epsilon_*^2 = \frac{\sum_i^M \sum_j^M \beta_i \beta_j \epsilon_i \epsilon_j (f_i, f_j)}{\sum_i^M \sum_j^M \beta_i \beta_j (f_i, f_j)} . \qquad (4\text{-}63)$$

In matrix notation, we write

$$\epsilon_*^2 = \frac{\beta^T E F E \beta}{\beta^T F \beta} \qquad (4\text{-}64)$$

where

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_M \end{bmatrix} \qquad E = \begin{bmatrix} \epsilon_1 & & & 0 \\ & \epsilon_2 & & \\ & & \ddots & \\ 0 & & & \epsilon_M \end{bmatrix} \qquad (4\text{-}65)$$

and

$$F = \begin{bmatrix} (f_1, f_1) & (f_1, f_2) & \cdots & (f_1, f_M) \\ \vdots & & & \\ (f_M, f_1) & (f_M, f_2) & \cdots & (f_M, f_M) \end{bmatrix} \qquad (4\text{-}66)$$

F is positive definite since its determinant is a Grams determinant. We want to find the maximum error incurred in approximating a sum of exponentials by elements of the subspace spanned by the functions $\epsilon^{-a_1 t}, \ldots \epsilon^{-a_N t}$. This maximum error determines the dimensionality of the set of function

$$F^{M}_{\alpha_i} = \left\{ f(t):\ f(t) = \sum_{i=1}^{M} \beta_i \, \epsilon^{-\alpha_i t},\ \beta_i \text{ real numbers} \right\}$$

with respect to the set of functions $\epsilon^{-a_1 t}, \ldots \epsilon^{-a_N t}$, where dimensionality is as defined in Eq. (4-5). We seek then the maximum of

$$\frac{\beta^T E F E \beta}{\beta^T F \beta} . \tag{4-68}$$

A theorem by Courant [42] p. 66, states that the maximum of

$$\frac{x^T A x}{x^T P x}$$

is given the largest eigenvalue of the matrix

$$P^{-1} A \tag{4-69}$$

where P is definite .

In our case then, we seek the largest eigenvalue of

$$F^{-1} E F E. \tag{4-70}$$

This largest error may not be attained for a particular ensemble, as the β_i may not take on all possible values. Eq. (4-70) then represents an upper bound on the representation error. Practically, it would probably be more convenient to compute the actual errors using Eq. (4-45). This problem could be pursued farther by seeking

$$\min_{a_i,\ i = 1, N} \{ \text{largest eigenvalue of } F^{-1} E F E \} \tag{4-71}$$

but as was remarked earlier, the approximation error is relatively insensitive to small perturbations of the a_i, and also it is shown in the following discussion that such a minimal representation is not really needed. The

goodness of the approximation for a given N may be estimated using Eq. (4-25) where the f's are given by Eq. (4-62). The largest eigenvalue of Equation (4-70) represents the largest normalized error of approximating an element of the subspace spanned by $\epsilon^{-\alpha_1 t}, \ldots, \epsilon^{-\alpha_M t}$ by an element of the subspace spanned by $\epsilon^{-a_1 t}, \ldots, \epsilon^{-a_N t}$.

Example

To illustrate the simplicity of computation afforded using the derived error expressions, we consider the approximation of $f_1 = \epsilon^{-t}$, $f_2 = \epsilon^{-3t}$, $f_3 = \epsilon^{-5t}$ by the subspace spanned by $g_1 = \epsilon^{-2t}$ and $g_1 = \epsilon^{-4t}$. Of interest are the normalized approximation errors and the comparison of (f_i, f_j) and $(\hat{f}_i, \hat{f}_j)$. Using the expressions derived above we find immediately that

$$\epsilon_1 = \left(\frac{1-2}{1+2}\right)\left(\frac{1-4}{1+4}\right) = +\frac{1}{5}, \quad \epsilon_1^2 = \frac{1}{25}$$

$$\epsilon_2 = \left(\frac{3-2}{3+2}\right)\left(\frac{3-4}{3+4}\right) = -\frac{1}{35}, \quad \epsilon_2^2 = \frac{1}{1225}$$

$$\epsilon_3 = \left(\frac{5-2}{5+2}\right)\left(\frac{5-4}{5+4}\right) = +\frac{1}{21}, \quad \epsilon_3^2 = \frac{1}{441}$$

and

$$(\hat{f}_1,\hat{f}_2) = (f_1,f_2)\ [1 - \epsilon_1 \epsilon_2] = \frac{1}{4}\left[1 + \frac{1}{(5)(35)}\right] = \frac{44}{175}$$

$$(\hat{f}_1,\hat{f}_3) = (f_1,f_3)\ [1 - \epsilon_1 \epsilon_3] = \frac{1}{6}\left[1 - \frac{1}{(5)(21)}\right] = \frac{52}{(15)(21)}$$

$$(\hat{f}_2,\hat{f}_3) = (f_2,f_3)\ [1 - \epsilon_2 \epsilon_3] = \frac{1}{8}\left[1 + \frac{1}{(35)(21)}\right] = \frac{92}{(35)(21)}.$$

For comparison, we compute ϵ_1^2 using

$$\epsilon_1^2 = \frac{1}{||f_1||^2} \frac{G[f_1, g_1, g_2]}{G[g_1, g_2]}$$

$$= \frac{1}{||f_1||^2} \frac{\begin{vmatrix} (f_1,f_1) & (f_1,g_1) & (f_1,g_2) \\ (g_1,f_1) & (g_1,g_1) & (g_1,g_2) \\ (g_2,f_1) & (g_2,g_1) & (g_2,g_2) \end{vmatrix}}{\begin{vmatrix} (g_1,g_1) & (g_1,g_2) \\ (g_2,g_1) & (g_2,g_2) \end{vmatrix}}$$

$$= \left(\frac{1}{2}\right) \frac{\begin{vmatrix} \frac{1}{2} & \frac{1}{3} & \frac{1}{5} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{6} \\ \frac{1}{5} & \frac{1}{6} & \frac{1}{8} \end{vmatrix}}{\begin{vmatrix} \frac{1}{4} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{8} \end{vmatrix}}$$

Expanding the above determinants and dividing, yields finally

$$\epsilon_1^2 = 2\left(\frac{1}{50}\right) = \frac{1}{25}$$

and it is seen that there is more labor involved in finding only one of the normalizal errors using this procedure than finding all three of the errors and the three $(\hat{f}_i, \hat{f}_j)$ using the first technique. Moreover, $(\hat{f}_i, \hat{f}_j)$ cannot be found using this latter method.

The quantities ϵ_i^2, $(\hat{f}_i, \hat{f}_j)$ may also be found by orthonormalizing the set of function g_1, g_2 yielding ψ_1, ψ_2 and computing

$$\epsilon_i^2 = \frac{||f_i||^2 - \sum_j (f_i, \psi_j)^2}{||f_i||^2}$$

and $f_i = (f_i, \psi_1)\,\psi_1 + (f_i, \psi_2)\,\psi_2$. One must not only find ψ_1 and ψ_2, but the f_i's must be found in order to compute (f_i, f_j). For comparison, we compute (f_1, f_2). The ψ's may be found by orthonormalizing the set of function g_1, g_2 using the Gram-Schmidt procedure or the equivalent Kartz procedure. One set of ψ's is

$$\psi_1 = 2\epsilon^{-2t}$$

$$\psi_2 = \sqrt{8}[-2\epsilon^{-2t} + 3\epsilon^{-4t}].$$

Then $(f_1, \psi_1) = \int_0^\infty 2\epsilon^{-2t}\,\epsilon^{-t} dt = 2/3$

$$(f_1, \psi_2) = \int_0^\infty \epsilon^{-t}\sqrt{8}[-2\epsilon^{-2t} + 3\epsilon^{-4t}]\,dt = \sqrt{8}\left[\frac{-1}{15}\right]$$

$$(f_2, \psi_1) = \int_0^\infty 2\epsilon^{-3t}\,\epsilon^{-2t}\,dt = \frac{2}{5}$$

$$(f_2, \psi_2) = \int_0^\infty \epsilon^{-3t}\sqrt{8}\,[-2\epsilon^{-2t} + 3\epsilon^{-4t}]\,dt = \sqrt{8}\left(\frac{1}{35}\right).$$

$$\hat{f}_1 = \frac{2}{3}\psi_1 + \sqrt{8}\left(\frac{-1}{15}\right)\psi_2$$

$$\hat{f}_2 = \frac{2}{5}\psi_1 + \sqrt{8}\left(\frac{1}{35}\right)\psi_2$$

finally,

$(\hat{f}_1, \hat{f}_2) = \frac{4}{15} - \frac{8}{(15)(35)} = \frac{44}{175}$. This result agrees with that found using the first method, but was obtained with considerably more effort.

If now the exponents of g_1 and g_2 are changed in an attempt to obtain a better representation, the latter two methods require a complete re-computation, while the first method allows the desired quantities to be found with a minimum of effort.

4.11 Signal Estimation

As was remarked earlier in this chapter, several authors have made use of the idea of characterizing a received signal x by its projections onto a known set of signals Φ^N. If a sufficient number of the φ_i are used, and the φ_i are complete in the space from which x is taken, then x can be characterized to within an arbitrarily small degree of accuracy. If the signal x was known exactly, the optimum receiver would be a filter matched to x. If $x_N(t) = \sum^N a_i\varphi_i(t)$ is an approximation to x relative to the φ_i, and the receiver is a filter matched to x_N, then the system is subject to the errors considered in Chapter II. Even if x is N-dimensional relative to the φ_i's, the measurements of its coordinates are invariably corrupted by noise so that an exact characterization of even a finite-dimensional signal is not possible and one can only make estimates of the coordinates. If we assume that N is sufficiently large so that $x_N(t)$ approximates any possible received signal x(t) within a prescribed tolerable error, then the problem becomes that of estimating the a_i in the expression $x_N(t) = \sum^N a_i\varphi_i(t)$. The estimate of $x_N(t)$ is denoted by $\hat{x}_N(t) = \sum^N \hat{a}_i\varphi_i(t)$ where the $\hat{a}_i$'s are the estimates of the a_i's. The received waveform is considered to be of the form

$$y(t) = n(t) + x_N(t)$$

where n(t) is the additive noise. The simplest possible estimate of a_k is to take

$$a_k = \int_0^T y(t)\,\varphi_k(t)\,dt = \int_0^T u(t)\,\varphi_j(t) + \int_0^T \Big(\sum^N a_i\varphi_i(t)\Big)\varphi_k(t)$$

$$= n_k + a_k \,.$$

If the noise is white and gaussian, this estimate is a conditional maximum likelihood estimate. For a detailed treatment of estimation of linear signal parameters see Glazer [15] and Parks [31]. If the noise is white zero mean with spectral density N_o, we have that

$$E[n_k] = 0$$

and $$E[n_k^2] = N_o$$

since the φ_k are orthonormal. In other words, the same amount of noise corrupts the measurement of each signal coordinate. If the signals to be estimated have the same energy

$$E = \int_0^T x_N^2(t) = \sum^N a_i^2$$

then, in general, the larger the required N, the smaller (on the average) are the a_i, and since the measurement noise remains the same, one would like to use the smallest N consistent with the allowable error. For any finite N, however, we may take a sufficient number (M) of observations and make the effect of the noise arbitrarily small. This may be done by taking the sample mean

$$\hat{a}_{k_M} = \frac{1}{M} \sum_{j=1}^{M} \hat{a}_{k_j}$$

$$= a_k + \frac{1}{M} \sum_{j=1}^{M} n_{k_j} \qquad (4\text{-}72)$$

$$= a_k + n_{k_M}.$$

Since the noise is white zero mean and the φ_i are orthonormal, the n_{k_j} are uncorrelated and

$$E[u_{k_M}] = 0$$
$$E[u_{k_M}^2] = N_o/M. \qquad (4\text{-}73)$$

By taking M sufficiently large the variance associated with the estimate of an individual coordinate may be made arbitrarily small. However, the quantity of interest is the error of approximating $\hat{x}_N$ where we define

$$\epsilon = \frac{\int [x_N(t) - \hat{x}_N(t)]^2}{\int [x_N(t)]^2}$$

$$= \frac{\sum_{}^{N} n_i^2}{\sum a_i^2} = \frac{\sum_{}^{N} n_i^2}{E}$$

where $E = \int x_N^2(t) = \sum_{}^{N} a_i^2$ is the total signal energy. Without loss of generality we take $N_o = 1$. E then represents the signal-noise ratio. Write $z = \sum_{}^{N} n_i^2$. If the noise is gaussian, z is the sum of the squares of N independent zero mean, unit variance gaussian random variables. The probability

density function $p(z)$ is given by the well-known chi-square density function

$$p(z) = \frac{1}{2\left(\frac{N}{2}\right)\Gamma\left(\frac{N}{2}\right)} z^{\frac{N}{2}-1} \epsilon^{-\frac{z}{2}} .$$

Recall that N is the required number of signal coordinates. Since the error ϵ is always greater than zero, a reasonable criterion is to choose a tolerable error γ and decide on a confidence level P where P is the probability that $\epsilon \leq \gamma$ or $z \leq \gamma E$. That is,

$$P = \int_0^{\gamma E} p(z)\, dz.$$

If M observations are made and an estimate based on the M observations are made according to Eq. (4-72), it is seen from Eq. (4-73) that the effect of making M observations is to increase the effective signal-noise ratio by a factor of M. That is, the probability that after making M observations the normalized approximation error is less than or equal to γ is given by

$$P_M = \int_0^{M\gamma E} p(z)\, dz$$

$$P_M = \int_0^{M\gamma E} \frac{1}{2^{\frac{N}{2}}\Gamma\frac{N}{2}} z^{\frac{N}{2}-1} \epsilon^{-\frac{z}{2}}\, dz .$$

This function is well tabulated. For instance the "Tables of the Incomplete Gamma-Function" edited by Karl Pearson may be used if in his $I(\mu,p)$ we set $p = \frac{N}{2} - 1$ and $M\gamma E = \mu\sqrt{2N}$, then

$$I(\mu,p) = P_M .$$

Recall that N is the number of coordinates required for a given error using a particular set of Φ^N. The number N will be different for different choices of

representation functions. For a particular class of signals, choosing the representation functions to be Laguerre function may require N = 20 while the same approximation error might be achieved by using only 5 sinusoids. In figure (4- 2) we plot P vs $M\gamma E$ with N as the parameter, and in figure (4-3) we plot $M\gamma E$ vs N with P as the parameter. The significance of being able to reduce the number os signal coordinates by proper choice of representation functions depends to some extent on the nature of the problem. If one may make any number of observation, then the signal may be estimated within an arbitrarily small error for any value of signal-noise ratio. However, the characteristics of the channel may be changing at such a rate that the signal is "essentially the same" for only a small number of observation intervals. In this case, in order to meet the allowable error requirement with a prescribed confidence level, one can either increase the signal-noise ratio (supply more transmitter power) or attempt to reduce the required number of signal coordinates. The choice of (the allowable error) depends upon the type of signaling scheme. Figures (2-6), (2-7), (2-8) in Chapter II indicate that if one is using one signal (the on-off case) an approximation error of 0.1 may be satisfactory, while if the system attempts to utilize two orthogonal signals, γ must be an order of magnitude smaller to ensure satisfactory performance. For P = 0.95 (the probability is 0.95 that the approximation error is within the allowable error requirement) the reduction of the required number of signal coordinates from say 20 to 10 allows a reduction of the quantity $M\gamma E$ from about 31.5 to about 18.3, a factor of 1.72. This is equivalent to increasing the signal-noise ratio by a factor of 1.72 for fixed M and γ, or if γ and E are fixed this reduction in the number of required signal coordinates allows the number of observations to be decreased by a factor of 1.72. Depending upon the signal-noise ratio, this reduction in M may or may not be significant. If the signal-

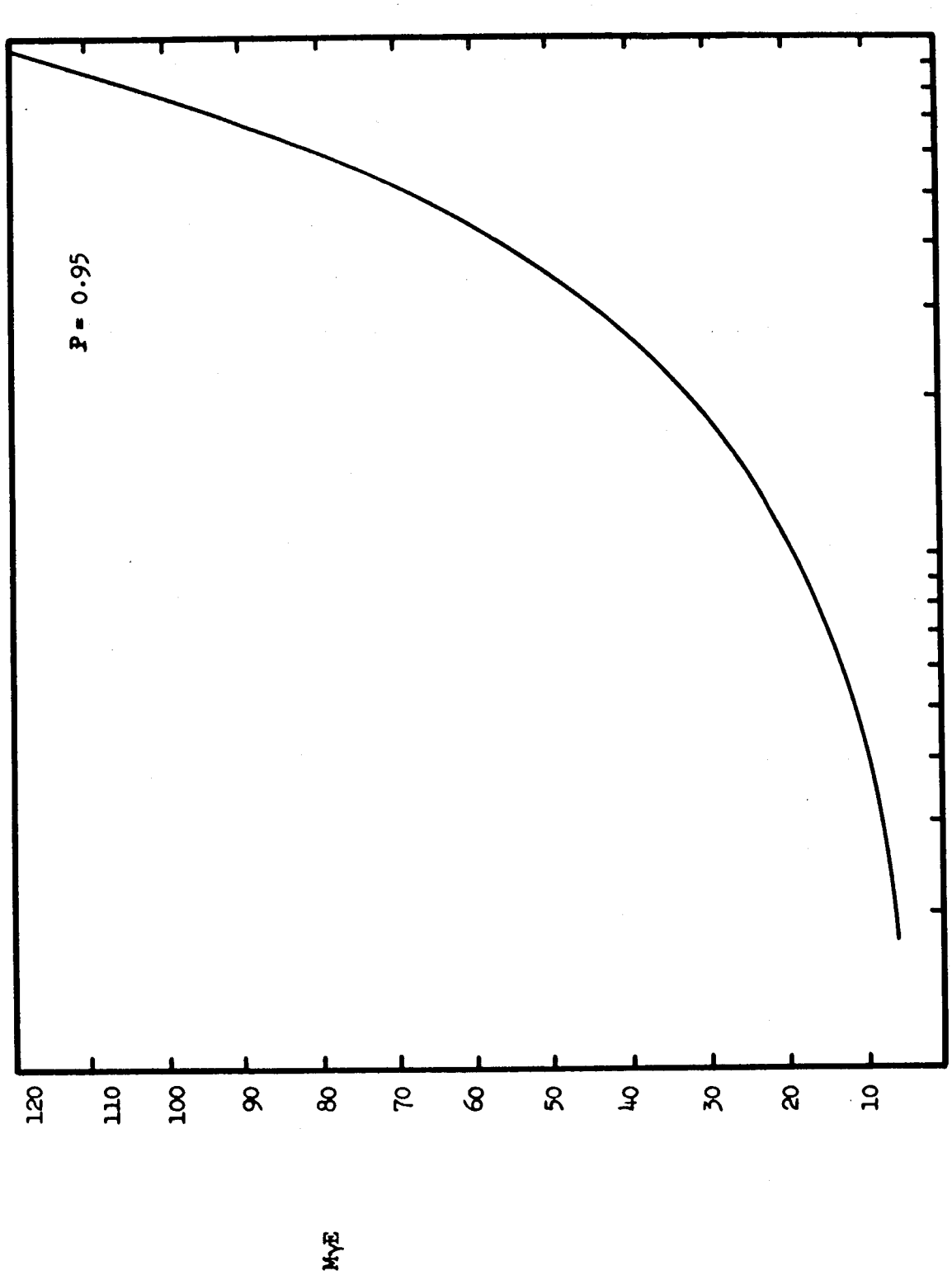


Fig. (4-3)

MyÆ versus N for P=0.95

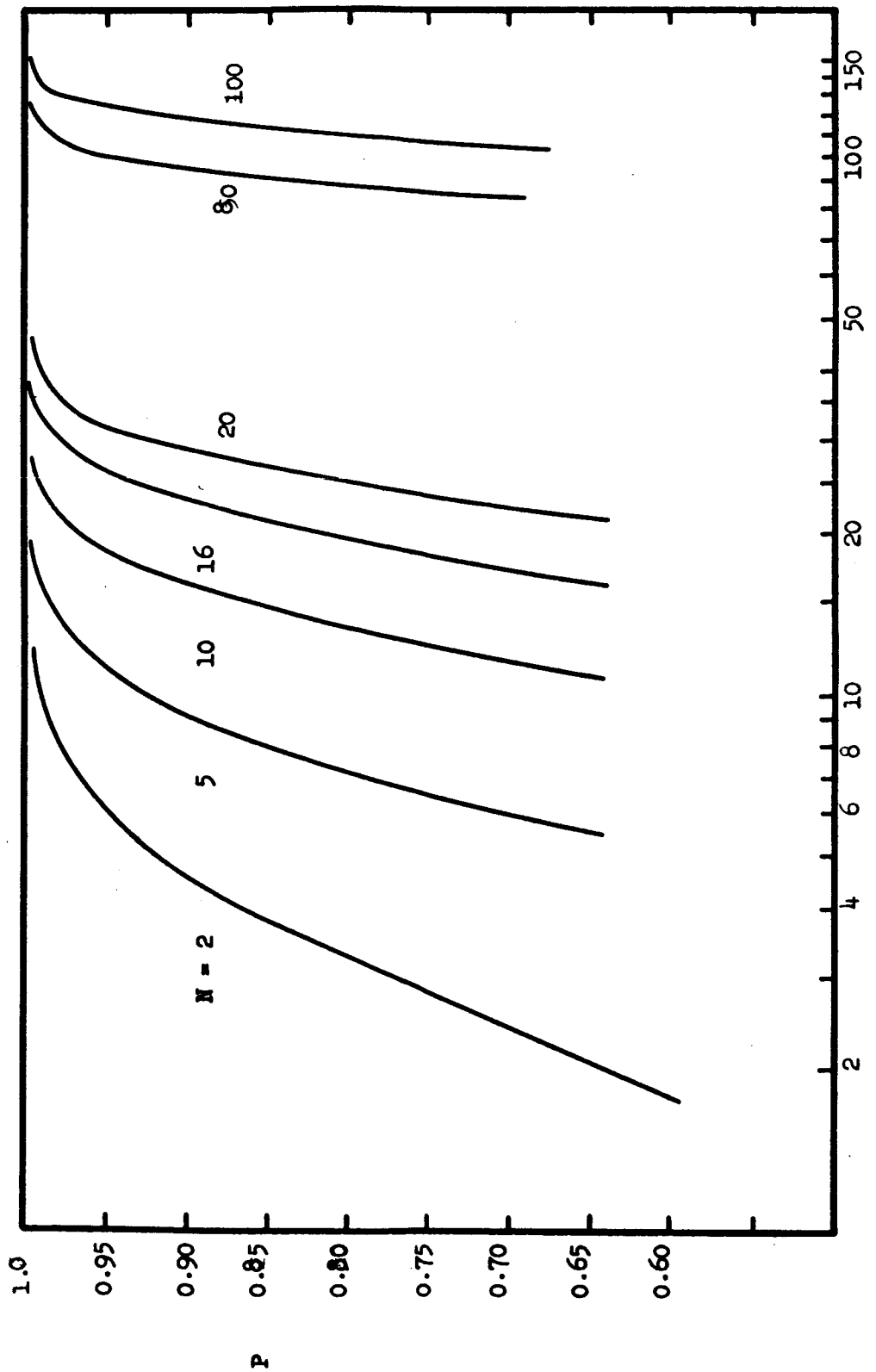


Fig. (4-2)

MγE

Confidence Level versus MγE for a Given Number of Signal Coordinates.

noise ratio is high, M may already be small (say 5) and reducing the number of observation to 3 by a better choice of signal coordinates is probably not worthwhile. However, if the signal-noise ratio is low, the reduction of M from say 32 to 18 (a factor of 1.72) is significant. It is seen from figures (4-2) and (4-3) that a search for a minimal number of signal coordinates is in general not really needed. For example, if the allowable error requirement could be met with a minimum N = 16 rather than N = 20, the quantity $M\gamma E$ is decreased only by a factor of about 1.1, and for the same signal-noise ratio as in the previous case, would mean a reduction of M from 32 to 30; a decrease hardly worth the effort of seeking any better signal coordinates. This bears out the statement made earlier in connection with exponential function approximation that a minimal representation was not really needed. However it is seen from the above that a poor choice of signal coordinates (i.e. many more than are necessary) can materially affect system performance.

The techniques of Chapter III for constructing transmitted signals to produce desirable received signals depended upon being able to measure the inner produce $\left(\int_T \theta_i(t)\ \theta_i(t)\ dt \right)$ of the channel's responses to the generating signals. The inner products can be measured in any manner. In general this can be done by approximating θ_i by

$$\theta_i = \sum_{k=1}^{N} a_{ik}\ \psi_k$$

$$(\theta_i, \theta_j) = \sum_{j}^{N} a_{ik}\ a_{jk}$$

If the set of ψ_k is complete, by taking N sufficiently large the above approximation can be made as close to equality as is desired. If the channel characteristics do not change with time, then of course an arbitrarily large amount of time may be spent making measurements to an arbitrary degree of

accuracy using any finite number of signals coordinates. In this case, the problem of optimum or even "good" choices of signal coordinates is not a factor in the system design.

Actually the channel characteristics do change with time and it is this phenomena that "adaptive" systems are supposed to combat. However, to the authors knowledge, there has been no analysis made of a system where the changing characteristics of the channel is taken into account. That is, the system "adapts" to a fixed but unknown signal (or signal parameters). Even if the signal is considered to be a sample function from a random process, once it is selected it is fixed. It is merely assumed that if the characteristics vary "slowly" enough the system will "track" the slowly changing signal. It is only when the measurement must be made in a given number of observations (before the channel characteristics change "appreciably") that consideration must be given to the selection of signal coordinates.

Summary

In this chapter an attempt was made to clarify and make more precise the concepts associated with "best representation functions and niminal finite-dimensional signal representation. It was seen that the intuitive notion of a set of signals being "approximately finite dimensional" raises rather deep mathematical problems when the statement is made more precise. The selection of a best or minimal set of representation functions for an arbitrary ensemble of signals is in general not possible. Not the least of the difficulties associated with such a problem is in adequatley describing the ensemble. In an actual physical situation, a complete description, and hence a truly minimal representation is precluded since one is able to observe only a finite number of the members of the ensemble. One can only assume that the observed members of the ensemble are "representative" of the ensemble in the

sense that if a set of approximating function is constructed to approximate the observed members of the ensemble, the approximation is also acceptable for the other members of the ensemble. In this connection, methods were developed for obtaining an estimate of the goodness of any N-dimensional set of approximating functions based on the observation of M members of the ensemble.

If an ensemble of signals is initially characterized by its projections onto a set of M exponential functions $\epsilon^{-\alpha_1 t}, \ldots, \epsilon^{-\alpha_M t}$ where M is taken sufficiently large to ensure adequate characterization, the dimensionality of the representation may be reduced by approximating signals of the form

$$f = \beta_1 \epsilon^{-\alpha_1 t} + \ldots + \beta_M \epsilon^{-\alpha_M t}$$

by functions of the form $f = \lambda_1 \epsilon^{-a_1 t} + \ldots \lambda_N \epsilon^{-a_N t}$. In this connection, very simple and computationally useful expressions were developed for the error

$$\epsilon^2 = || f - \hat{f} ||^2$$

The selection of the signal coordinates (or the representation functions) in connection with "adaptive" communications problems is seen to be a relevant factor when observation or measurement time is limited, although a minimal representation is not required.

Chapter 5

CONCLUDING REMARKS

The selection of optimum signals and the attending approximation error has been based on the assumption that the optimual receiver was a correlator or matched filter. If the statistics of the noise is gaussian, the optimum receiver does have the form of a matched filter. If the noise is not gaussian, the form of the optimum receiver is not known except for a few special cases (e.g. Rayleigh noise). This lack of knowledge of the form of the optimum receiver for non-gaussian noise precludes the selection of optimum signals for these cases. The error due to filter and signal mismatch computed in Chapter II is valid for coherent detection or arbitrary signals. The extreme sensitivity of systems designed to receiver orthogonal signals is more likely to occur when the signals occupy the same time interval and the same bandwidth. FSK, for example, suffers only slightly from mismatch error when the two frequencies are sufficiently for apart. A study of some interst would be the sensitivity of systems designed to receive more than two orthogonal signals. The technique used here for the binary case should prove useful for such an investigation.

For channels having bandwidths comparable to those of the transmitted signals, considerable waveform distortion may be present. This waveform distortion provides one source of mismatch error considered in Chapter II. In Chapter III, computationally simple techniques were developed for constructing transmitted signals so that the received signals would, for example, be orthogonal. The receiver may be a discrete receiver, distinguishing

between the received signals by using, for example, a number of time samples of the waveforms. However, the transmitter must transmit through the channel analog waveform, (i.e. not a sequence of numbers), so that when the signals pass through a waveform-distorting channel, the signal construction must be bases on analog waveforms even though the receiver operates only on the time samples. The procedure given here is valid for coherent reception. That is, the signals are constructed to have a prescribed inner product matrix for a given observation interval. If the observation interval is slightly different from the interval over which the signals were designed to be, say, orthogonal, the resulting inner product matrix may be different from the desired one. That is, one would also like the autocorrelation and cross-correlation functions to be such that imperfect synchronization does not materially affect the performance of the system. A study combining the techniques developed in Chapter III with the correlation properties of the signals would be of considerable interest.

The problem of finding "best" finite-dimensional subspace for representing signals from a given class was seen to be revalent to the detection problem if the number of coordinates is fixed, and the signals are known to be of a particular class. The number of coordinates required to yield a good representation of the signal has no effect on the performance of the system. That is, if $N\psi$'s or $M\psi$'s are required to represent the signals, both produce the same test statistic. If however, the number of coordinates is not sufficient to represent the signal, the discrete receiver is subject to the type of error considered in Chapter II. In adaptive receivers where estimates are made of the received waveforms, the selection of signal coordinates enters the problem somewhat differently. Here the number of observation intervals required to form a good estimate of the signal waveform depends upon the number of signal coordinates required. It was shown that

for the important case of conditional maximum likelihood estimates the number of required coordinates is not critical, although the number of required signal coordinates plays an important ade when the channel characteristics are changing in such a manner that the signal may be assumed to be essentially the same for only a small number of observations.

However, actually finding "best" finite-dimensional representations for given classes of signals, where the problem is precisely stated, is at best difficult, and in any practical situation probably impossible. The problem has received comparatively little attention from mathematicions with apparent good reason. The relative merits of any set of approximating signals must be judged in a particular application. In certain applications exponential functions have proven to provide satisfactory approximations with a remarkably small number of exponentials. Aside from their usefulness in particular applications, exponential functions possess many useful properties as demonstrated in the references. The error expressions derived in Chapter IV further extends the usefulness of exponential function approximation. In this connection, an investigation of other parametric families of functions having approximation properties similar to those of exponentials may prove to be of significant value in theoretical investigations.

BIBLIOGRAPHY

BIBLIOGRAPHY

1. W. Davenport and W. Root, Random Signals and Noise, McGraw-Hill Book Co., Inc., 1958.

2. D. Middleton, An Introduction to Statistical Communication Theory, McGraw-Hill Book Co., Inc., 1961.

3. S. G. Mikhlin, Integral Equations, Pergamon Press, N.Y., 1957.

4. J. H. H. Chalk, "The Optimum Pulse Shapte for Pulse Communication," Proc. I.E.E., March 1950.

5. R. Weinstock, Calculus of Vanations, International Series in Pure and Applied Mathematics, McGraw-Hill Book Co., Inc., 1952.

6. I. Gerst and J. Diamond, "The Elimination of Intersymbol Interference by Input Signal Shoping," Proc. I.R.E., July 1961.

7. J. C. Hancock, H. Schwarzlander, R.E. Totty, "Optimization of Pulse Transmission," Proc. I. R.E., October 1962.

8. N. I. Acheiser, Theory of Approximation, Ungar Publishing Co., N.Y., 1956.

9. Baghdady ed., Lectures on Communication System Theory, McGraw-Hill Book Co., Inc., 1960.

10. A. V. Boldkrishnan, "A Contribution to the Sphere-Packing Problem of Communication Theory," J. Math. Anal. and Appl., December 1961.

11. A. H. Nuttall, "Error Probabilities for Equi-Correlated M-ary Signals Under Phase-Coherent and Phase-Incoherent Reception," I.R.E. Trnas. on Inf. Theory, July 1962.

12. D. Slepian, "Report on Progress in Information Theory in the U.S.A. 1960-1963," I.E.E.E. Trans. on Inf. Theory, October 1963.

13. C. A. Stutt, "Information Rate in a Continuous Channel for Regular Simplex Codes," I.R. E. Trans. on Inf. Theory, December 1960.

14. E. A. Guillemin, The Mathematics of Circuir Analysis, M.I.T. Technology Press, 1949.

15. E. M. Glazer, "Signal Detection by Adaptive Filters," I.R.E. Trans. on Inf. Theory, April 1961.

16. C. V. Jakowatz, et. al., "Adaptive Waveform Recognition," Proc. Fourth London Symp. on Inf. Theory.

17. D. Jackson, The Theory of Approximation, Am. Math. Soc. Coll. Publ. XI, New York, 1930.

18. P. Koyovkin, Linear Operators and Approximation Theory, Delhi 1960.

19. M. Golomb, Lectures on Theory of Approximation, Argonne National Laboratory, 1963.

20. E. A. Guillemin, "What is Natures Error Criterion", I.R.E. Trans. on Circuit theory, March 1954.

21. J. L. Brown, "Mean Square Truncation Error in Series Expansions of Random Functions," J. Siam Vol. 8, March 1960.

22. A. Koschmann, "On the Filtering of Nonstationary Time Series," Doctoral Dissertation, School of Electrical Engineering, Purdue University, August 1954.

23. D. Slepian, H. Landau, H. Pollack, "Prolate Spheriiodal Wave Functions, Fourier Analysis and Uncertainty," Bell Telephone System Monograph 3746.

24. H. Landau and H. Pollack, "Prolate Spheriodal Wavefunctions, Fourier Analysis and Uncertainty - III," Bell Telephone System Monograph 4238.

25. W. H. Huggins,"Representation and Analysis of Signals; the Use of Orthogonalized Exponentials," The Johns Hopkins University School of Electrical Engineering, Reports Number AFRC TN-58-191.

26. D. Lai, "An Orthonormal Filter for Exponential Waveforms," the Johns Hopkins University School of Electrical Engineering, Report Number AFRC TN-58-191.

27. T. Y. Young and W. H. Huggins, "On the Representation of Electrocardiograms", I.E.E.E. Trans. on Bio-medical Electronics, July 1963.

28. R. McDonough, "Matched Exponents for the Representation of Signals," The Johns Hopkins University School of Electrical Engineering, Report.

29. W. H. Kautz, "Transient Synthesis in the Time Domain," I.R.E. Trnas. on Circuit Theory, September 1954.

30. R. Covrant and D. Hilbert, Methods of Mathematical Physics, Vol. I, Interscience Publisheres, Inc., New York, 1953.

31. J. Parks, "Statistical Estimation of Normalized Signal Parameters", The Johns Hopkins University Radiation Laboratory, Tech. Rept. AF-72.

32. E. Bodewig, Matrix Calculus, Interscience Publishers, Inc., New York 1959.

33. H. Schwarzlander, "Certain Optimum Signaling Waveforms for Channels with Memory," Doctoral Dissertation, School of Electrical Engineering, Purdue University, August 1964.

34. W. H. Huggins, "Signal Theory," I.R.E. Trans. on Circuit Theory, December 1956.

35. C. W. Helstrom, "Statistical Theory of Signal Detection," Pergamon Press, Oxford, 1960.

APPENDIX

APPENDIX

The following definitions are taken from Chapter I of [8].

[A1] The Concept of Metric Space. A set E having the elements $x,y,z,\ldots$ is known as a metric space, and the elements are called points of the space, if for every pair of elements x,y there can be found a corresponding non-negative number $D[x,y]$ which is called the distance between the points x and y, and which satisfies the following conditions:

A. $D[x,x] = 0$

B. $D[x,y] = D[y,x] > 0$ (if $x \neq y$)

C. $D[x,z] \leq D[x,y] + D[y,z]$ (triangular inequality)

[A2] The Concept of Linear Normalized Space. A set E having the elements $x,y,z,\ldots$ is called a linear normalized space, the elements themselves points, vectors, or functions, if

1. There is defined in E an operation, which we called addition and denote by the symbol +, in respect to which E forms an abelian group; the zero element of the group E will be denoted by 0;

2. A multiplication of the elements of the set E by (real or complex) numbers α, β, , ... is defined so that

$$\alpha(x+y) = \alpha x+\alpha y$$
$$(\alpha+\beta)x = \alpha x+\beta y$$
$$\alpha(\beta x) = (\alpha\beta)x$$
$$1.\ x = x$$
$$0.\ x = 0$$

3. To every element $x \in E$ there corresponds a certain positive number $||x||$ called the norm of the element x, which satisfies the conditions

$$||x|| = 0 \text{ if and only if } x = 0$$

$$||\alpha x|| = |\alpha| \, ||x||$$

$$||x+y|| \leq ||x|| + ||y||$$

[A3] The space L^p $(p \geq 1)$. By $L^p[a,b]$ is meant the totality of all functions measureable in the interval [a,b], whose absolute value to the p^{th} power is integrable in the sense of Leberque. In this connection, addition and multiplication with complex numbers are to be considered in the ordinary sense. Two elements $x = x(t)$, $y = y(t) \in L^p$ are identified if the equality $x(t) = y(t)$ holds almost everywhere.

The norm is defined by

$$||x|| = \left\{ \int_a^b |x(t)|^p \right\}^{\frac{1}{p}}$$

It can be shown that L^p is a linear normalized space. For $p = 2$, L^2 denotes the space of function with finite energy. Without any practical loss of generality, the function may be considered to be piecewise continuous and the integral taken in the normal Riemann sense.

[A4] Uniqueness of the Approximation. The expression $\lambda_1 g_1 + \ldots + \lambda_n g_n$ which furnishes the best approximations of the element x is uniquely determined when the space is strictly normalized, i.e. if the equality sign in the inequality

$$||x+y|| \leq ||x|| + ||y|| \qquad (x \neq 0, \; y \neq 0)$$

holds only for $y = \alpha x$ $(\alpha \geq 0)$.

The space $L^{\rho}(\rho > 1)$ furnishes an example of a strictly normalized space, while the space of continuous functions on [a,b] with $||x||_c = \max_{a \leq t \leq b} |x(t)|$ is not strictly normalized.

[A5] Hilbert Space. Achieser defines Hilbert space as a linear space (i.e. the first two conditions of [A2] are fulfilled) in which for every pair of function x and y there is a corresponding number (x,y) (we restrict ourselves here to real quantaties) called the scalar product (or inner product) of the functions x and y, and satisfying the following conditions.

a) $(y,x) = (x,y)$

b) $(\alpha_1 x_1 + \alpha_2 x_2, y) = \alpha_1 (x,y) + \alpha_2 (x_2,y)$

c) $(x,x) \geq 0$

d) $(x,x) = 0$ if and only if $x = 0$

Hilbert space represents a linear normalized if we put

$$||x|| = (x,x)^{\frac{1}{2}} .$$

The space L^2 represents an example of a Hilbert space if we put $(x,y) = \int_a^b x(t)\, y(t)\, dt$.